Visual Cortex and Deep Networks

LEARNING INVARIANT REPRESENTATIONS

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Visual Cortex and Deep Networks

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Computational Neuroscience
Terence J. Sejnowski and Tomaso A. Poggio, editors

For a complete list of books in this series, see the back of the book and http://mitpress.mit.edu/books/series/computational-neuroscience

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Visual Cortex and Deep Networks
Learning Invariant Representations

Tomaso A. Poggio
Fabio Anselmi

The MIT Press
Cambridge, Massachusetts
London, England

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

© 2016 Massachusetts Institute of Technology

This work is subject to a Creative Commons CC-BY-NC-ND license.

Subject to such license, all rights are reserved.

Funding for the open access edition was provided by the MIT Libraries Open Monograph Fund.

Library of Congress Cataloging-in-Publication Data

Names: Poggio, Tomaso, author. | Anselmi, Fabio, author.
Title: Visual cortex and deep networks : learning invariant representations / Tomaso A. Poggio and Fabio Anselmi.
Description: Cambridge, MA : MIT Press, [2016] | Series: Computational neuroscience | Includes bibliographical references and index.
Identifiers: LCCN 2016005774 | ISBN 9780262034722 (hardcover : alk. paper)
Subjects: LCSH: Visual cortex. | Vision. | Neural networks (Neurobiology) | Perceptual learning. | Computational neuroscience.
Classification: LCC QP383.15 .P64 2016 | DDC 612.8—dc23 LC record available at http://lccn.loc.gov/2016005774

10 9 8 7 6 5 4 3 2 1

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Contents

Series Foreword ix
Preface xi

1 Invariant Representations: Mathematics of Invariance 1
1.1 Introduction and Motivation 1
1.2 Invariance Reduces Sample Complexity of Learning 4
1.3 Unsupervised Learning and Computation of an Invariant Signature (One-Layer Architecture) 5
1.4 Partially Observable Groups 9
1.5 Optimal Templates for Scale and Position Invariance Are Gabor Functions 10
1.6 Quasi Invariance to Nongroup Transformations Requires Class-Specific Templates 11
1.7 Two Stages in the Computation of an Invariant Signature: Extension of the HW Module to Hierarchical Architectures 14
1.8 Deep Networks and i-Theory 21

2 Biophysical Mechanisms of Invariance: Unsupervised Learning, Tuning, and Pooling 23
2.1 A Single-Cell Model of Simple and Complex Cells 23
2.2 Learning the Wiring in the Single-Cell Model 24
2.3 Hebb Synapses and Principal Components 24
2.4 Spectral Theory and Pooling 26
2.5 Tuning of Simple Cells 28

3 Retinotopic Areas: V1, V2, V4 29
3.1 V1 29
3.1.1 A New Model from i-Theory for Eccentricity Dependence of Receptive Fields 29

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

3.1.2 Fovea and Foveola 32
3.1.3 Scale and Position Invariance in V1 34
3.1.4 Tuning of Cells in V1 35
3.2 V2 and V4 35
3.2.1 Multistage Pooling 35
3.2.2 Predictions of Crowding Properties in the Foveola and Outside It (Bouma's Law) 36
3.2.3 Scale and Shift Invariance Dictates the Architecture of the Retina and the Retinotopic Cortex 39
3.2.4 Tuning of Simple Cells in V2 and V4 42

4 Class-Specific Approximate Invariance in Inferior Temporal Cortex 45
4.1 From Generic Templates to Class-Specific Tuning 45
4.2 Development of Class-Specific and Object-Specific Modules 45
4.3 Domain-Specific Regions in the Ventral Stream 48
4.4 Tuning in the Inferior Temporal Cortex 49
4.5 Mirror-Symmetric Tuning in the Face Patches and Pooling over Principal Components 49

5 Discussion 53
5.1 i-Theory: Main Ideas 53
5.2 i-Theory, Deep Learning Networks, and the Visual Cortex 54
5.3 Predictions and Explanations 55
5.4 Remarks 56
5.5 Ideas for Research 61

Appendix 65
A.1 Invariant Representations and Bounds on Learning Rates: Sample Complexity 65
A.1.1 Translation Group 66
A.1.2 Scale and Translation: 1-D Affine Group 67
A.2 One-Layer Architecture: Invariance and Selectivity 68
A.2.1 Equivalence between Orbits and Probability Distributions 68
A.2.2 Cramér-Wold Theorem and Random Projections for Probability Distributions 69
A.2.3 Finite Random Projections Almost Discriminate among Different Probability Distributions 70
A.2.4 Number of Templates Depends on Pooling Size 72

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A.2.5 Partially Observable Groups 75
A.2.6 Approximate Invariance to Nongroup Transformations 79
A.3 Multilayer Architecture: Invariance, Covariance, and Selectivity 80
A.3.1 Recursive Definition of Simple and Complex Responses 80
A.3.2 Inheriting Transformations: Covariance 82
A.4 Complex Cells: Wiring and Invariance 86
A.5 Mirror-Symmetric Templates Lead to Odd-Even Covariance Eigenfunctions 86
A.6 Gabor-like Shapes from Translation Group 87
A.6.1 Spectral Properties of Template Transformations Covariance Operator: Cortical Equation 87
A.6.2 Cortical Equation: Derivation and Solution 89
A.7 Gabor-like Wavelets 92
A.7.1 Derivation from an Invariance Argument for 1-D Affine Group 92
A.7.2 Wavelet Templates from Best Invariance and Heisenberg Principle 94
A.8 Transformations with Lie Group Local Structure: HW Module 97
A.9 Factorization of Invariances 98
A.10 Invariant Representations and Their Cost in Number of Orbit Elements 101
A.10.1 Discriminability Cost for Hierarchical versus Single-Layer Architecture 102
A.10.2 Invariance Cost for Hierarchical versus Single-Layer architecture 104
A.11 Nonlinearities Are Key in Gaining Compression from Repeated Random Projections 106
A.11.1 Hierarchical (Linear) Johnson-Lindenstrauss Lemma 106
A.11.2 Example of Nonlinear Extension of Johnson-Lindenstrauss Lemma 107
References 109
Index 117

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Series Foreword

Computational neuroscience is an approach to understanding the development and function of nervous systems at many different structural scales, including the biophysical, the circuit, and the system levels. Methods include theoretical analysis and modeling of neurons, networks, and brain systems and are complementary to empirical techniques in neuroscience. Areas and topics of particular interest to this book series include computational mechanisms in neurons, analysis of signal processing in neural circuits, representation of sensory information, system models of learning and memory. Further topics of interest include the intersection of computational neuroscience with engineering, from representation and dynamics, to observation and control.

Terrence J. Sejnowski
Tomaso Poggio

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Preface

The ventral visual stream is believed to underlie object recognition abilities in primates. Fifty years of modeling efforts, which started with the original Hubel-Wiesel proposal of a hierarchical architecture iterating in different layers the motif of simple and complex cells in V1, led to a series of quantitative models [2, 3] that are increasingly faithful to the biological architecture and are able to mimic properties of cells in different visual areas while achieving humanlike recognition performance under restricted conditions. Deep convolutional learning networks, which are hierarchical and similar to the HMAX model but otherwise do not respect the ventral stream architecture and physiology, have been trained with very large labeled data sets [4]. The population of resulting model neurons mimics well the object recognition performance of the macaque visual cortex (DiCarlo). However, the nature of the computations carried out in the ventral stream is not explained by such models, which can be simulated on a computer but remain otherwise rather opaque.

In this book we develop a mathematical framework describing learning of invariant representations in the ventral stream. Our theory, called i-theory, applies to a broad class of hierarchical networks that pool over transformations. In particular, it applies to deep convolutional learning networks where the transformations are just translations in $\mathbb{R}^2$. The networks associated with the theory do not have nonlinearities other than for pooling. In particular, i-theory applies to networks with pooling nonlinearities (to compute a histogram or associated statistics) such as sigmoidal threshold units or linear rectifiers (see figure 1 and chapter 1).

The following chapters outline a proposal and a theory for the feedforward computation of invariant representations in the ventral stream, that is, a theory of the first 100 msec of visual perception, from the onset of an image to activation of inferior temporal (IT) cortex neurons about 100 msec later. In particular, such representations are likely to underlie rapid categorization, that is, immediate object recognition from flashed images [5, 6]. We emphasize that

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

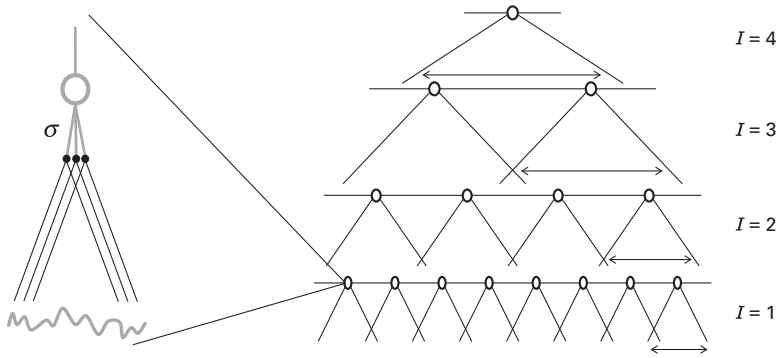

Figure 1

A hierarchy of Hubel-Wiesel modules. In each module *(inset)* the black dots represent simple cells calculating dot products of images with transformed templates (implementing a convolution) and filtering with, for example, a sigmoid function with different thresholds. The outputs of the simple cells are summed by a complex cell (pooling). This set of operations is indicated by the symbol $\bigwedge$.

the theory is not a full theory of vision that also will have to explain the top-down effect and the role of backprojections, but only part of it. The theory is based on the hypothesis that the main computational goal of the ventral stream is to compute neural representations of images that are invariant to transformations commonly encountered in the visual environment and learned from unsupervised experience. i-theory proposes computational explanations for various aspects of the ventral stream architecture and of its neurons. It makes several testable predictions. It also leads to network implementations that show high performance in object recognition benchmarks [7]. As mentioned, the theory is based on the unsupervised, automatic learning of invariant representations. Since invariant representations turn out to be good representations for supervised learning, characterized by small sample complexity, the architecture of the ventral stream may ultimately be dictated by the need to learn, at least during the initial phase of development, from very few labeled examples. We use i-theory to summarize several key aspects of the neuroscience of visual recognition, explain them, and predict others. The book is organized into four chapters, a discussion, and an appendix.

Chapter 1 describes the general theoretical framework of a computational theory of invariance. In particular, we describe relevant new mathematical results on invariant representations in vision; the appendix provides details and proofs (see also [1, 8, 10]). The starting point is a theorem proving that image

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

representations that are invariant to translation and scaling and approximately invariant to some other transformations (e.g., face pose and expressions) can considerably reduce the sample complexity of learning. We then describe how an invariant and unique (selective) signature can be computed for each image or image patch. The invariance can be exact in the case of locally compact group transformations (we focus on groups such as the affine group in 2-D and one of its subgroups, the similitude group consisting of translation and uniform scaling) and approximate under nongroup transformations. A module performing filtering and pooling, like the simple and complex cells described by the Hubel-Wiesel (HW) module, can compute such estimates. Each HW module provides a feature vector, which we call a signature, for the part of the visual field that is inside its receptive field. Interestingly, Gabor functions turn out to be optimal templates for maximizing simultaneous invariance to translation and scale. Hierarchies of HW modules inherit their properties, while alleviating the problem of clutter in the recognition of wholes and parts. Finally, the same HW modules at high levels in the hierarchy are shown to be able to compute representations that are approximately invariant to a much broader range of transformations, such as 3-D expression of a face, pose of a body, and viewpoint, by using templates (reflected in the neuron's tuning) that are highly specific for each object class.

Chapter 2 describes how neuronal circuits may implement the operation required by the HW algorithm. It specifically discusses new models of simple and complex cells in V1. It also introduces plausible biophysical mechanisms for tuning and pooling and for learning the wiring based on Hebbian-like unsupervised learning. The rest of the book is devoted to reviewing the application of the theory to the feedforward path of the ventral stream in the primate visual cortex.

Chapter 3 applies the theory to explain the multiresolution, eccentricity-dependent architecture of the retina and V1 as a consequence of the need for simultaneous space and scale invariance. It predicts several still untested properties of the early stages of vision. Chapter 3 also deals with V2 and V4 as higher layers in the hierarchy devoted to progressively increase invariance to shift and scaling while minimizing interference from clutter.

Chapter 4 is about the final IT stage, where class-specific representations that are quasi-invariant to nongeneric transformations are computed from a shift and scale invariant representation obtained from V4. It also discusses the modular organization of anterior IT in terms of the theory. In particular, it proposes an explanation of the architecture and of some puzzling properties of the face patches system.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Chapter 5 discusses predictions to be tested and other open problems.

The book integrates work on invariance in object recognition done in our group since 2011. The results have appeared in several different technical reports and papers available online, for which we are indebted to our collaborators Lorenzo Rosasco, Joel Leibo, Jim Mutch, Andrea Tacchetti, Qianli Liao, Leyla Isik, and Cheston Tan (see [1, 7–9, 11–15]). This is the first time we attempt to organize them in book form. The book is intended to address a broad audience from neuroscientists to mathematicians. Readers not interested in mathematical details can skip the more formal appendix without compromising understanding of the main text.

Further developments, mentioned in chapter 1 (end of section 1.7) are already under way. They will be fully explored in a planned sequel to this book.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

1 Invariant Representations

Mathematics of Invariance

1.1 Introduction and Motivation

One could argue that the most important aspect of intelligence is the ability to learn [16]. How do present supervised learning algorithms compare with brains? One of the most obvious differences is the ability of people and animals to learn from very few labeled examples. A child, or a monkey, can learn a recognition task from just a few examples. The main motivation of this book is the conjecture that the key to reducing the sample complexity of object recognition is invariance to transformations. Images of the same object usually differ from each other because of simple transformations such as translation, scale (distance), or more complex deformations such as viewpoint (rotation in depth) or change in pose (of a body) or expression (of a face).

The conjecture is supported by previous theoretical work showing that *almost all the complexity* in recognition tasks is due to the viewpoint and illumination nuisances that swamp the intrinsic characteristics of the object [17]. It implies that in many cases recognition, both identification (e.g., of a specific car relative to other cars) and categorization (e.g., distinguishing between cars and airplanes) would be much easier (only a small number of training examples would be needed to achieve a given level of performance) if the images of objects were rectified with respect to all transformations, or equivalently, *if the image representation itself were invariant*.

The case of identification is obvious, since the difficulty in recognizing exactly the same object like an individual face is only due to transformations. In the case of categorization, consider the suggestive evidence from the classification task in figure 1.1. The figure shows that if an oracle factors out all transformations in images of many different cars and airplanes, providing rectified images with respect to viewpoint, illumination, position, and scale, the problem of categorizing cars versus airplanes becomes easy; it can be done accurately with very few labeled examples. In this case, good performance

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

was obtained from a single training image of each class, using a simple classifier. In other words, the sample complexity of the problem seems to be very low (see box 1.1 and figure 1.1).

A similar argument involves estimating the cardinality of the universe of possible images generated by different viewpoints, such as variations in scale, position, and rotation in 3-D, versus true intraclass variability, such as different types of cars. Let us try to estimate whether the cardinality of the universe of possible images generated by an object originates more from intraclass variability (e.g., different types of dogs) or from the range of possible viewpoints, including scale, position, and rotation in 3-D. Assuming a granularity of a few minutes of arc in terms of resolution and a visual field of, say, 10 degrees, one would get 10^3–10^5 different images of the same object from x, y translations, another factor of 10^3–10^5 from rotations in depth, a factor of 10–10^2 from rotations in the image plane, and another factor of 10–10^2 from scaling. This gives on the order of 10^8–10^{14} distinguishable images for a single object. On the other hand, how many different distinguishable (for humans) types of dogs exist within the dog category? It is unlikely that there are more than, say, 10^2–10^3. From this point of view, it is a much greater win to be able to factor out the geometric transformations than the intracategory differences.

As context for this book, let us describe the conceptual framework for primate vision that we use:

- The first 100 msec of vision in the ventral stream are mostly feedforward. The main computation goal is to generate a number of image representations, each one quasi-invariant to some transformations experienced during development, such as scaling, translation, and pose changes. The representations are used to answer basic default questions about what kind of image and what may be there.
- The answers will often have low confidence, requiring an additional verification/prediction step, which may require a sequence of shifts of gaze and attentional changes. This step may rely on generative models and probabilistic inference or on top-down visual routines following memory access. Routines that can be synthesized on demand as a function of the visual task are needed in any case to go beyond object classification. Note that in a Turing test of vision [133] only the simplest, standard questions (what is there? who is there? etc.) can be answered by pretrained classifiers.

We consider only the feedforward architecture of the ventral stream and its computational function. To help readers understand more easily the mathematics of this part of the book, we anticipate here the network of visual areas that we propose for computing invariant representations for feedforward visual

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

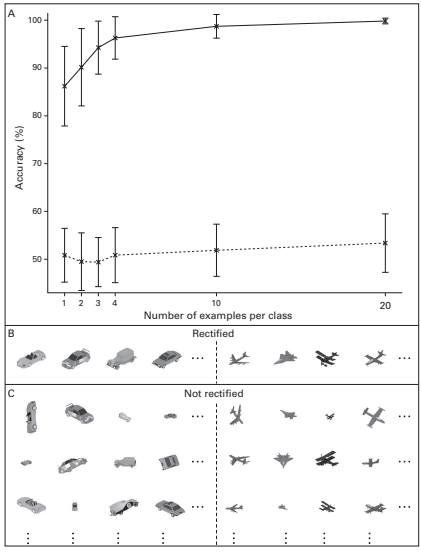

Figure 1.1

Sample complexity for the task of categorizing cars versus airplanes from their raw pixel representations (no preprocessing). (A) Performance of a nearest-neighbor classifier (distance metric = 1 − correlation) as a function of the number of examples per class used for training. Each test used 74 randomly chosen images to evaluate the classifier. Error bars represent ± 1 standard deviation computed over 100 training/testing splits using different images out of the full set of 440 objects × number of transformation conditions. *Solid line:* the rectified task. Classifier performance for the case where all training and test images are rectified with respect to all transformations; example images shown in (B). *Dashed line:* the unrectified task. Classifier performance for the case where variation in position, scale, direction of illumination, and rotation around any axis (including rotation in depth) is allowed; example images shown in (C). The images were created using 3-D models from the Digimation model bank and rendered with Blender.

recognition. There are two main stages. The first one comprises retinotopic areas computing a representation that is invariant to affine transformations. The second computes approximate invariance to object-specific, nongroup transformations. The second stage consists of parallel pathways, each one for a different object class (see figure 1.5, stage 1). The results of this part do not strictly require these two stages. If both are present, as it seems is the case for the primate ventral stream, the mathematics of the theory requires that the object-specific stage follow the one dealing with affine transformations. According to the i-theory, the Hubel-Wised module is the basic module for both stages. The first- and second-stage pathways may each consist of a single layer of HW modules. However, mitigation of interference by clutter requires a hierarchy of layers (possibly corresponding to visual areas such as V1, V2, V4, PIT) within the first stage. It may not be required in visual systems with lower

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

resolution, such as in the mouse. The final architecture is shown in figure 1.5. In the first stage about four layers compute representations that are increasingly invariant to translation and scale, while in the second stage a large number of specific parallel pathways deal with approximate invariance to transformations that are specific for objects and object classes. Note that for any representation that is invariant to X and selective for Y, there may be a dual representation that is invariant to Y but selective for X. In general, they are both needed for different tasks, and both can be computed by an HW module with different pooling strategies. In general, the circuits computing them share a good deal of overlap. For example, it is possible that different face patches in the cortex are used to represent different combinations of invariance and selectivity.

1.2 Invariance Reduces Sample Complexity of Learning

In a machine learning context, invariance to image translations, for instance, can be built up trivially by memorizing examples of the specific object in different positions. Human vision, on the other hand, is clearly invariant for novel objects: people do not have any problem in recognizing in a distance-invariant way a face seen only once. It is intuitive that representations of images that are invariant to transformations such as scaling, illumination, and pose should allow supervised learning from far fewer examples.

A proof of the conjecture for the special case of translation or scale or rotation is provided in appendix section A.1. For images defined on a grid of pixels, the result (in the case of translations) can be proved using well-known relations between covering numbers and sample complexity.

Box 1.1
Sample Complexity

Sample complexity is the number of examples needed for the estimate of a target function to be within a given error rate. In the example of figure1.1, the number of airplanes or cars, we trained the linear classifier to perform the recognition task with a certain precision.

Sample Complexity for Translation Invariance

Consider a space of images of dimensions $p \times p$, which may appear in any position within a window of size $rp \times rp$. The natural image representation yields a sample complexity (for a linear classifier) of order $m_{\text{image}} = O(r^2p^2)$; the invariant representation yields a sample complexity of order

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

$$m_{\text{inv}} = O(p^2). \tag{1.1}$$

This simple observation says that in the case of a translation group, an invariant representation can decrease considerably the sample complexity, that is, the number of supervised examples necessary for a certain level of accuracy in classification. A heuristic rule is that the sample complexity gain is on the order of the number of virtual examples generated by the action of the group on a single image (see [18, 19]). This is not a constructive result, but it supports the hypothesis that the ventral stream in the visual cortex tries to approximate such an oracle. The next section describes a biologically plausible algorithm that the ventral stream may use to implement an invariant representation.

1.3 Unsupervised Learning and Computation of an Invariant Signature (One-Layer Architecture)

The following algorithm is biologically plausible, as we discuss in detail in chapter 2, where we argue that it may be implemented in the cortex by an HW module, that is, a set of *KH* complex cells (see box 1.2) with the same receptive field, each pooling the output of a set of simple cells whose sets of synaptic weights correspond to the K templates of the algorithm and its transformations (which are also called templates) and whose output is filtered by a sigmoid function with Δh threshold, $h = 1, \cdots, H,\ \Delta > 0$.

Box 1.2

Simple Cells, Complex Cells [20]

A threshold vector product can be interpreted as the output of a neuron, called *simple cell*, which computes a possibly high-dimensional inner product with a template t and applies a nonlinear operation to it. In this interpretation, eq. (1.2) can be seen as the output of the pooling of many simple cells by a second neuron *complex cell*, which aggregates the output of other neurons by a simple averaging operation. Neurons of the former kind can be found in the visual cortex.

The algorithm for groups (finite or compact, defined on a finite or compact set) (see figure 1.2) is as follows.

Developmental stage

1. For each of K isolated (on an empty background) objects, or templates, memorize a sequence Λ of $|G|$ frames corresponding to the object's transformations $(g_i, i = 1, \cdots, |G|)$. For now, we suppose the g_i to belong to a finite group G; see box 1.3. The sequence of frames is observed over a time

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

interval; thus $\Lambda = \{g_0 t, g_1 t, \cdots, g_{|G|} t\}$ for template t. (For template t^k the corresponding sequence of transformations is denoted Λ_k.)

2. Repeat for each of K templates.

Runtime computation of invariant signature for a single image of any new object

1. For each t^k compute the dot product of the image with each of the $|G|$ transformations in Λ_k.
2. For each k compute the cumulative histogram of the resulting values.
3. The signature is the set of K cumulative histograms, that is, the set of

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\left\langle I, g_i t^k \right\rangle + h\Delta), \tag{1.2}$$

where I is an image, σ is a threshold function, $\Delta > 0$ is the width of bin in the histogram, and $h = 1, \cdots, H$ is the index of the bins of the histogram.

Box 1.3
Group Transformations

A group is a set of objects equipped with a law of composition, $*$, such that

- a composition of any two elements of the set gives another element of the set (closure);
- the composition is associative, i.e., $(a * b) * c = a * (b * c)$ (associativity);
- the set includes an identity element $a * e = a$ (identity);
- the set includes all the inverses of the elements, a^{-1} (inverse).

The algorithm consists of two parts. The first part is unsupervised learning of transformations by storing transformed templates, which are images. This part is possibly done only once during development of the visual system. The second part is the actual computation of invariant signatures during visual perception. Our analysis is not restricted to the case of group transformations. For now, we consider groups that are compact and, for simplicity, finite.

This algorithm we used throughout the book. The guarantees we can provide depend on the type of transformations. The main questions are whether the signature is invariant under the same type of transformations that were observed in the first stage, and whether the signature is selective, for instance, can it distinguish between N different objects. In summary, as shown in appendix section A.2, the HW algorithm is invariant and selective (when K

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

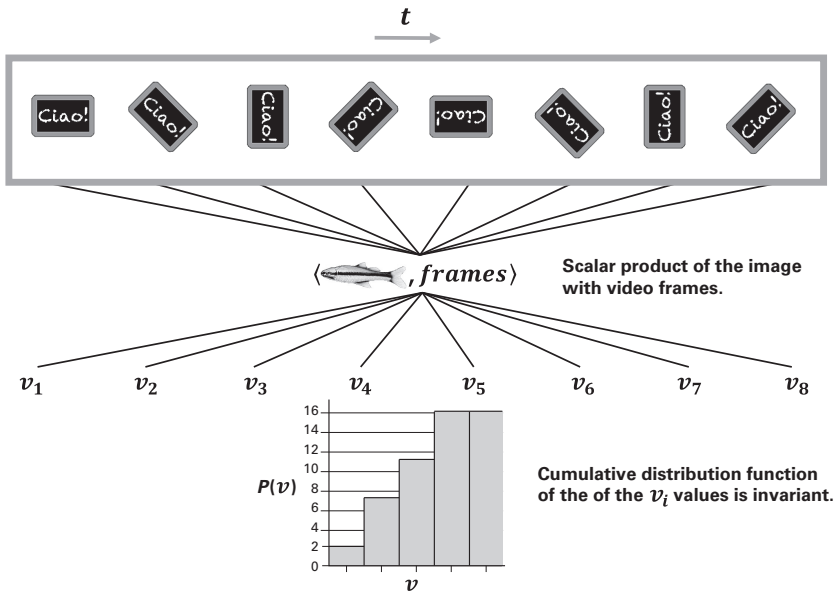

Figure 1.2

A graphical summary of the HW algorithm. The set of $\mu_h^k(I) = 1/|G| \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + h\Delta)$ values (see eq. (1.2)) corresponds to the the histogram where, e.g., $k = 1$ denotes the template blackboard, h the bins of the histogram; transformations g_i are from the rotation group. Crucially, mechanisms capable of computing invariant representations under affine transformations can be learned (and maintained) in an unsupervised, automatic way by storing sets of transformed templates that are unrelated to the object to be represented in an invariant way. In particular, the templates could be random patterns.

is order $\log(N)$) if the transformations form a group (but see also appendix section A.2.4). In this case, any set of randomly chosen templates will work for the first stage. Seen as transformations from a 2-D image to a 2-D image, the natural choice is the affine group consisting of translations, rotations in the image plane, scaling (possibly nonisotropic), and compositions thereof. The HW algorithm can learn with exact invariance and desired selectivity in the case of the affine group or its subgroups. In the case of 3-D images consisting of voxels with x, y, z coordinates, rotations in 3-D are also a group and in principle can be dealt with, achieving exact invariance from generic templates by the HW algorithm (in practice, this is rarely possible because of correspondence problems and self-occlusions). In section 1.6 we show that the same HW algorithm provides approximate invariance (under some conditions) for non-group transformations such as the transformations from 3-D to 2-D induced by 3-D rotations of an object.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

In the case of compact groups, the guarantees of invariance and selectivity are provided by the following two theorems (given informally here; detailed formulation in appendix section A.2 and in [8, 10]).

Box 1.4
Invariance and Group Average

For any function $f \in L^2(\mathbb{R})$ and compact group G,

$$\bar{f}(x) = \int dg\, f(gx)$$

is invariant. This property can be simply proved using the fact that the group is closed under its composition rule and the invariance of the Haar measure dg:

$$\bar{f}(\tilde{g}x) = \int dg\, f(\tilde{g}gx) = \int dg\, f(gx) = \bar{f}(x), \quad \forall \tilde{g} \in G,$$

i.e., $\bar{f}$ is invariant to G.

Invariance Theorem

The distributions represented by eq. (1.2) are invariant, that is, each bin is invariant

$$\mu_h^k(I) = \mu_h^k(gI) \tag{1.3}$$

for any g in G, where G is the group of transformations labeled g_i in eq. (1.2). The proof is based on the fact that the signature is a group average (see box 1.4).

Selectivity Theorem

For groups of transformations (e.g., the affine group), the distributions represented by eq. (1.2) can achieve any desired selectivity for an image among N images in the sense that they can ϵ-approximate the true distance between each pair of the images (and any transform of them) with probability $1 - \delta$ provided that

$$K > \frac{c}{\epsilon^2} \ln \frac{N}{\delta}, \tag{1.4}$$

where c is a universal constant.

The signature provided by the K cumulative histograms is a feature vector corresponding to the activity of the (HK) complex cells associated with the HW module. It is selective in the sense that it corresponds uniquely to an image of a specific object independently from its transformation. The stability of the signature under noisy measurements remains an open problem. Because

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

of the restricted dynamic range of cortical cells, the number H of bins is likely to be small, probably around 2 or 3 [21]. Note that other, related representations are possible (see [22]). A cumulative distribution function (cdf) is fully represented by all its moments; often a few moments, such as the average or the variance (energy model of complex cells; see [23]) or the max,

$$\mu^k_{\text{av}}(I) = \frac{1}{|G|}\sum_{i=1}^{|G|}\left\langle I, g_i t^k \right\rangle,$$

$$\mu^k_{\text{energy}}(I) = \frac{1}{|G|}\sum_{i=1}^{|G|}\left\langle I, g_i t^k \right\rangle^2,$$

$$\mu^k_{\text{max}} = \max_{g_i \in G}\left\langle I, g_i t^k \right\rangle,$$

can in practice replace the cumulative distribution function. Any linear combination of the moments is also invariant, and a small number of linear combinations is likely to be sufficiently selective.

1.4 Partially Observable Groups

This section outlines invariance, uniqueness, and stability properties of the signature obtained in the case in which transformations of a group are observable only within a *window* on the orbit (figure 1.3). The term *partially observable groups* (POGs) emphasizes the properties of the group—in particular, associated invariants—as seen by an observer (e.g., a neuron) looking through a window at a part of the orbit. Therefore the window should not be thought of only as a spatial window but over ranges of transformation parameters, for example, a window over scale and spatial transformations. With this observation in mind, let G be a finite group and $G_0 \subseteq G$ a subset (G_0 is not usually a subgroup). The subset of transformations G_0 can be seen as the set of transformations that can be *observed* through a window on the orbit (i.e., the transformations that correspond to a part of the orbit). A *local* signature associated with the partial observation of G can be defined considering

$$\mu^k_h(I) = \frac{1}{|G_0|}\sum_{i=1}^{|G_0|}\eta_h\left(\left\langle I, g_i t^k \right\rangle\right) \tag{1.5}$$

where η_h is a set of nonlinear functions and $\Sigma_{G_0}(I) = (\mu^k_h(I))_{h,k}$. This definition can be generalized to any locally compact group considering

$$\mu^k_h(I) = \frac{1}{V_0}\int_{G_0}\eta_h\left(\left\langle I, gt^k \right\rangle\right)dg, \quad V_0 = \int_{G_0} dg. \tag{1.6}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

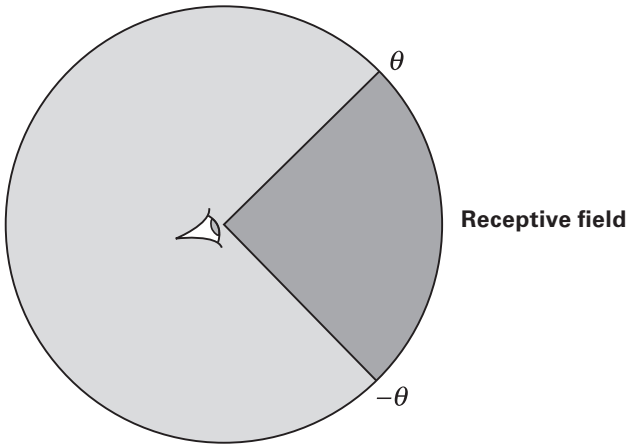

Figure 1.3

A partially observable compact shift: an object is seen through a window in different positions. The object is fully contained in the window and is isolated (blank background). A compact group of periodic translations acts on it; only a part of the orbit is observable.

The constant V_0 normalizes the Haar measure dg, restricted to G_0, so that it defines a probability distribution. The latter is the distribution of the images subject to the group transformations that are observable, that is, in G_0. These definitions can be compared to the definition in eq. (1.2) in the fully observable group case. We next discuss the properties of this signature. While uniqueness follows essentially from the analysis so far, invariance requires a new analysis.

1.5 Optimal Templates for Scale and Position Invariance Are Gabor Functions

The previous results apply to all groups, in particular to those that are not compact, but only locally compact, such as translation and scaling. In this case it can be proved that invariance holds within an observable window of transformations (see appendix section A.2.5 but also [1, 8]). For the standard HW module, the observable window corresponds to the receptive field of the complex cell (in space and scale). For maximum range of invariance within the observable window, it is proved (see appendix section A.2.5) that the templates must be maximally localized relative to generic input images. In the case of translation and scale invariance, this requirement is equivalent to localization in space and spatial frequency, respectively: templates must be maximally localized for maximum range of invariance in order to minimize

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

boundary effects due to the finite window. Assuming therefore that the templates are required to have simultaneously a minimum size in space and spatial frequency, it follows from results of Gabor [24, 25] that they must have a Gaussian envelope (for a certain definition of minimum spatial and frequency size). If the system of templates is required to be a frame then the following property holds.

Optimal Invariance Theorem

Gabor functions (here in $1-D$) of the form $t(x) = e^{-x^2/(2\sigma^2)}e^{i\omega_0 x}$ are the templates that give simultaneous maximal invariance for translation and scale (at each x and ω).

In general, templates chosen at random in the universe of images can provide scale and position invariance. However, for optimal invariance under scaling and translation, templates of the Gabor form are optimal (for a specific definition of *optimal*). This is the only computational justification we know of for the Gabor shape of simple cells in V1, which seems to be remarkably universal: it holds in primates [26], cats [27], and mice [28] (see figure 1.4 for results of simulations).

1.6 Quasi Invariance to Nongroup Transformations Requires Class-Specific Templates

All the previous results require a group structure and ensure exact invariance for a single new image. In 2-D all combinations of translation, scaling, and rotation in the image plane are included; transformations induced on the image plane by 3-D transformations, such as viewpoint changes and rotation in depth of an object, are not. The latter form a group in 3-D, that is, if images and templates were 3-D views. In principle, motion or stereopsis can provide the third dimensional though available psychophysical evidence [29, 30] suggests that human vision does not use it for recognition. Note that transformations in the image plane are affected not only by orthographic projection of the 3-D geometry but also by the process of image formation, which depends on the 3-D geometry of the object, its reflectance properties, and the relative location of light source and viewer.

It turns out that the same HW algorithm can still be applied to nongroup transformations such as transformations of the expression of a face or pose of a body, to provide under certain conditions approximate invariance around the center of such a transformation. In this case bounds on the invariance depend on specific details of the object and the transformation: we do not have general

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

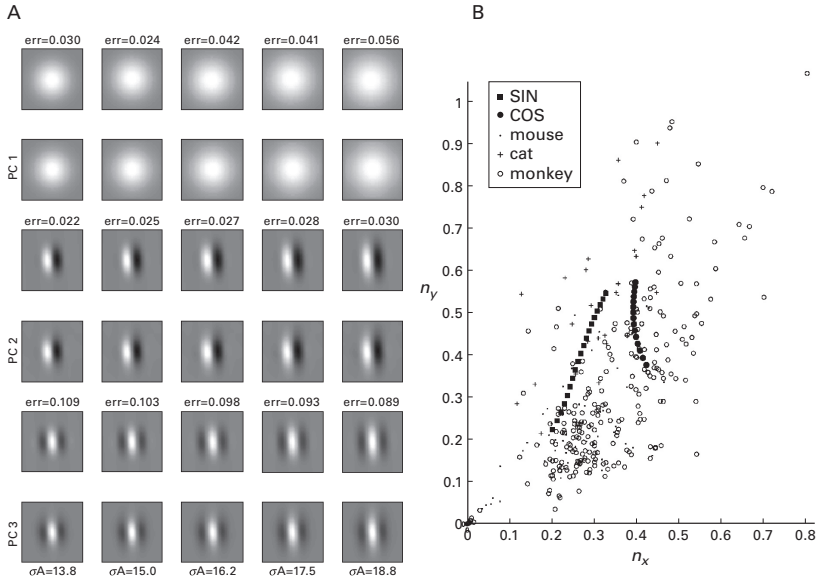

Figure 1.4

(A) Simulation results for V1 simple cells learning via principal component analysis. Each cell sees a set of images through a Gaussian window (its dendritic tree), shown in the top row. Each cell then learns the same weight vector, extracting the principal components of its input. (B) $n_y = \sigma_y/\lambda$ vs $n_x = \sigma_x/\lambda$ for the modulated (x) and unmodulated (y) direction of the Gabor wavelet. The slope σ_y/σ_x is a robust finding in the theory and apparently also in the physiology data. Neurophysiology data from monkeys, cats, and mice are reported together with our simulations. Source: [11].

results. The key technical requirement is that a condition of incoherence (see box 1.5) with respect to a transformation hold for the class of images I_C with respect to the dictionary t^k under the transformation T_r (we consider here a one-parameter transformation, r, for simplicity),

Box 1.5

Sparsity/Incoherence

A signal is sparse if its information content can be expressed with few coefficients when expanded in the right basis. For example, a sinusoid in the Fourier space is expressed by one coefficient, and therefore its representation is very sparse in the Fourier basis.

Two bases are called incoherent when signals are sparse when expressed in one basis and dense in the other. For a sinusoidal signal, the Fourier and time domain are incoherent. Incoherence extends the space-frequency duality to other bases.

$$\left\langle I_C, T_r t^k \right\rangle \approx 0, \quad |r| > a, \;\; a > 0. \tag{1.7}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Strictly speaking, the condition is valid when the object has localized support in the pooling region (is an isolated object). However, it holds approximately whenever the dot product has a fast decay with the transformation (e.g., wavelet coefficients). The property, which is an extension of the compressive sensing notion of mutual coherence, requires that the templates (an image can be considered a template) have a representation with sharply peaked correlation and autocorrelation (the constant a above in eq. (1.7) is related to the support of the peak of the correlation). Eq. (1.7) can be satisfied by templates that are similar to images in the set and are sufficiently rich to be incoherent for small transformations. Empirically, it appears, our incoherence condition is usually satisfied by the neural representation of images and templates at some high level of the hierarchy of HW modules. Like standard mutual incoherence (see [25]) our condition of incoherence with respect to a group is generic. Most neural patterns (templates and images from the same class) chosen at random will satisfy it. The full (theorem A.8 in appendix section A.2.5) takes the following form.

Class-Specific Property

$\mu_h^k(I)$ is approximatively invariant around a view if

- the dictionary of the templates relative to the transformations is incoherent,
- I is one of the templates and transforms in the same way as the templates;
- the transformation is smooth.

The main implication is that approximate invariance can be obtained for nongroup transformation by using templates specific to the class of objects. This means that class-specific modules are needed, one for each class; each module requires highly specific templates, that is, tuning of the cells. An example is face-tuned cells in the face patches. Unlike exact invariance for affine transformations, where tuning of the simple cells is nonspecific in the sense that does not depend on the type of image, nongroup transformations require highly tuned neurons and yield at best only approximate invariance.

Summary of the Results

The core of i-theory applies without qualification to compact groups such as rotations of the image in the image plane. Translation and scaling are, however, only locally compact. Each HW module usually observes only a part of the transformation's full range. Each module has a finite pooling range, corresponding to a finite window on the orbit associated with an image.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Exact invariance for each module is equivalent to a condition of *localization/sparsity* of the dot product between image and template. In the simple case of a group parameterized by one parameter r the condition is

$$\left\langle I, g_r t^k \right\rangle = 0 \quad |r| > a. \tag{1.8}$$

Since this condition is a form of sparsity of the generic image I with respect to a dictionary of templates t^k (under a group), this result may provide a justification for *sparse* encoding in sensory cortex (see, e.g., [31]). Localization yields the following surprising result: *optimal invariance for translation and scale implies Gabor functions as templates*. Since a frame of Gabor wavelets follows from natural requirements of completeness, this may also provide a general motivation for the scattering transform approach of Mallat based on wavelets [32]. Eq. (1.8), if relaxed to hold approximately, that is $\left\langle I_C, g_r t^k \right\rangle \approx 0 \quad |r| > a$, becomes a *sparsity condition for the class of I_C with respect to the dictionary t^k under the generic transformation T* (approximated locally by the group transformation g_r) when restricted to a subclass I_C of similar images. This property, which is similar to compressive sensing incoherence (but in a group context), requires that I and t^k have a representation with rather sharply peaked autocorrelation (and correlation). When the condition is satisfied, the basic module equipped with such templates can provide *approximate invariance* to nongroup transformations, such as rotations in depth of a face or its changes of expression (see appendix section A.2 and A.8). In summary, condition Eq. 1.8 can be satisfied in two different *regimes*. The first one, exact and valid for generic I, yields optimal Gabor templates. The second regime, approximate and valid for specific subclasses of I, yields highly tuned templates, specific for the subclass. This argument suggests generic, Gabor-like templates in the first layers of the hierarchy and highly specific templates at higher levels (incoherence improves with increasing dimensionality; see appendix section A.2.5).

1.7 Two Stages in the Computation of an Invariant Signature: Extension of the HW Module to Hierarchical Architectures

It is known that Hubel and Wiesel's original proposal [20] for visual area V1—of a module consisting of complex cells (C-units) combining the outputs of sets of simple cells (S-units) with identical orientation preferences but differing retinal positions—can be used to construct group-invariant detectors. This is the insight underlying many networks for visual recognition, including HMAX [33] and convolutional neural nets [2, 34]. We showed that a representation of images and image patches, in terms of a feature vector that is invariant to a broad range of transformations, such as translation, scale, expression of a

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

face, pose of a body, and viewpoint, makes it possible to recognize objects from only a few labeled examples. In the following we argue that hierarchical architectures of HW modules (indicated by $\bigwedge$ in figure 1.5) can provide such invariant representations while maintaining discriminative information about the original image. Each $\bigwedge$-module provides a feature vector (the signature) for the the part of the visual field that is inside its receptive field; the signature is invariant to ($\mathbb{R}^2$) affine transformations within the receptive field. The hierarchical architecture, since it computes a set of signatures for different parts of the image, is invariant to the more general family of locally affine transformations (which include globally affine transformations of the whole image). This reasoning also applies to invariance to global transformations that are not affine but are locally affine, that is, affine within the pooling range of some of the modules in the hierarchy (any differentiable transformation, no matter how complex, can be seen locally as an affine transformation). This remarkable invariance of the hierarchies follows from the key property of *covariance* (see box 1.6) of such architectures for certain image transformations and from the uniqueness and invariance of the individual module signatures. The basic HW module (see section 1.3) is at the core of the properties of the architecture.

Box 1.6
Covariance

In our theory, the key property of hierarchical architectures of layers of repeated HW modules—allowing the recursive use of single module properties at all layers—is the property of *covariance*: the response image at layer ℓ transforms like the response image at layer $\ell - 1$.

The main reasons for an extension to a hierarchical architecture like the are shown in figure 1.5 are the following:

- *Compositionality*. A hierarchical architecture provides signatures of larger and larger patches of the image in terms of lower-level signatures. Because of this, it can access memory in a way that matches naturally with the linguistic ability to describe a scene as a whole and as a hierarchy of parts.
- *Approximate factorization*. In architectures such as the network sketched in figure 1.5, approximate invariance to transformations specific for an object class can be learned and computed in different stages. This property may provide an advantage in terms of the sample complexity of multistage learning [16]. For instance, approximate class-specific invariance to pose (e.g., for faces) can be computed on top of a translation-and-scale-invariant

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

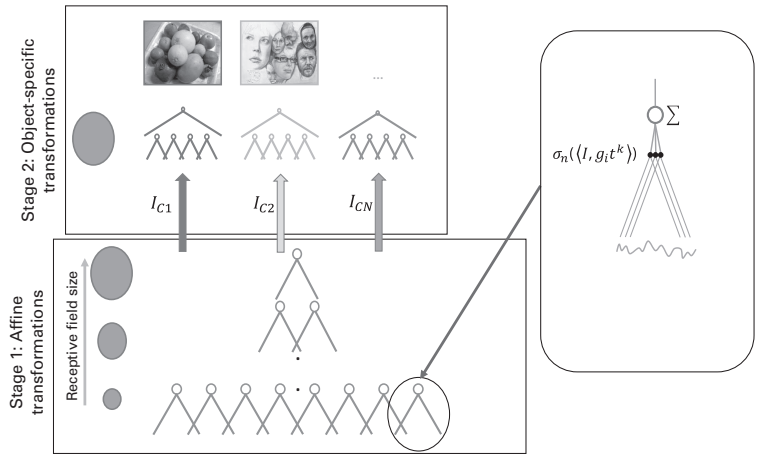

Figure 1.5

A hierarchical architecture of HW modules, indicated by $\bigwedge$. Signature provided by each of the nodes at each layer may be used by a supervised classifier. *Stage 1:* a hierarchy of HW modules *(inset)* with growing receptive fields provides a final signature (top of the hierarchy) that is globally invariant to affine transformations by pooling over a cascade of locally invariant signatures at each layer. *Stage 2:* transformation-specific modules provide invariance for nongroup transformations (e.g., rotation in depth).

representation [35]. Thus the implementation of invariance can, in some cases, be factorized into different steps corresponding to different transformations (see [36, 37] for related ideas).

- *Optimization of local connections* and optimal reuse of computational elements. Despite the high number of synapses on each neuron it would be impossible for a complex cell to pool information across all the simple cells needed to cover an entire image.
- *Minimization of number of transformations* per template needed to achieve invariance (see appendix section A.10).
- *Clutter tolerance* (see chapter 3).

Probably all these properties together are the reason that evolution developed hierarchies.

One-layer architectures are unable to capture the *hierarchical organization of the visual world*, where scenes are composed of objects that are themselves composed of parts. Objects (parts) can change position or scale in a scene relative to each other without changing their identity and often changing the scene only in a minor way. Thus global and local signatures from all levels of the hierarchy must be able to access memory in order to enable the categorization

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

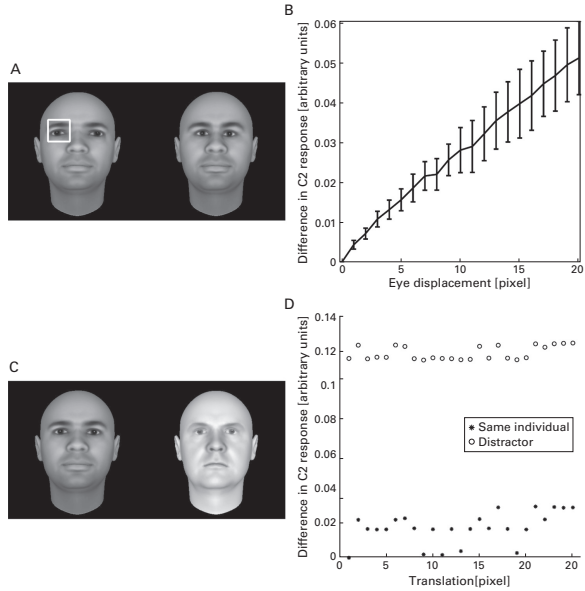

Figure 1.6

Empirical demonstration of the properties of invariance, stability, and uniqueness of the hierarchical architecture in a specific two-layer implementation (HMAX). (A) reference image on the left and a deformation of it (the eyes are closer to each other) on the right. (B) HW module at layer 2 (c_2) whose receptive fields contain the whole face provides a signature vector that is (Lipschitz) stable with respect to the deformation. In all cases, the figure shows just the Euclidean norm of the signature vector. The c_1 and c_2 vectors are not only invariant but also selective. Error bars represent ± 1 standard deviation. (C) two different images are presented at various locations in the visual field. (D) Euclidean distance between the signatures of a set of HW modules at layer 2 with the same receptive field (the whole image) and a reference vector. The signature vector is invariant to global translation and discriminative (between the two faces). In this example the HW module represents the top of a hierarchical, convolutional architecture. The images used were 200×200 pixels.

and identification of whole scenes as well as patches of the image corresponding to objects and their parts. Figure 1.6 shows examples of invariance and stability for wholes and parts. In the architecture of figure 1.5, each $\bigwedge$-module provides uniqueness, invariance, and stability at different levels, over increasing ranges from bottom to top. Thus, in addition to the desired properties of invariance, stability, and discriminability, these architectures match the hierarchical structure of the visual world and the need to retrieve items from memory at various levels of size and complexity.

The property of *compositionality* is related to the efficacy of hierarchical architectures versus one-layer architectures in dealing with the problem of

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

partial occlusion and the more difficult problem of clutter in object recognition. Hierarchical architectures are better at recognition in clutter than one-layer networks (see chapter 3 and [38]) because they provide signatures for image patches of several sizes and locations. However, hierarchical feedforward architectures cannot fully solve the problem of clutter. More complex (e.g., recurrent) architectures are likely needed for human-level recognition in clutter (see, e.g., [39–41]) and for other aspects of human vision.

A hierarchical architecture has layers with receptive fields of increasing size (figure 1.7). The intuition is that transformations represented at each level of the hierarchy begin with small affine transformations that is, over a small range of translation, scale, and rotation. The size of the transformations represented in the set of transformed templates will increase with the level of the hierarchy and the size of the apertures. In addition it seems intuitive that only translations will be seen by small apertures with scale and orientation changes being relevant later in the hierarchy.

To be more specific, suppose that the first layer consists of an array of small apertures (in fact corresponding to the receptive fields of V1 cells) and focus on one of the apertures. Box 1.7 explains why the only transformations that can

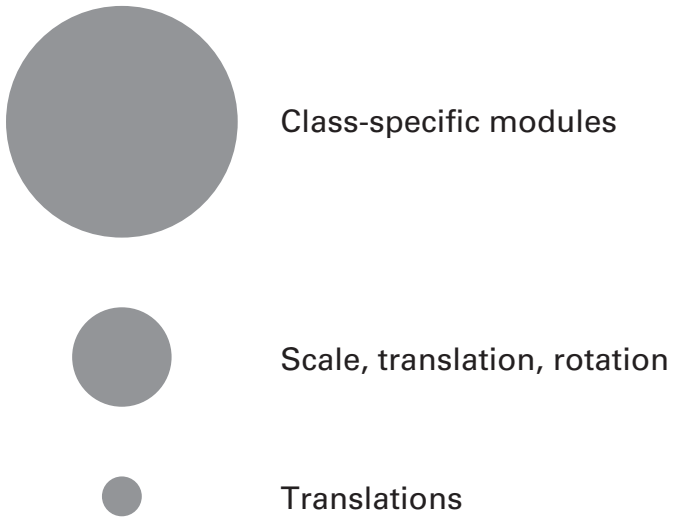

Figure 1.7

The conjecture is that receptive field sizes (of complex cells) affect not only the size but also the type of transformations that are learned and represented by the templates. In particular, small apertures (such as in complex cells in V1) only see small translations.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

be seen by a small aperture are small translations, even if the transformation of the image is more complex.

Box 1.7

Any transformation looked at through a small window is approximately an affine transformation.

Our approach differs in the assumption that small (close to identity) diffeomorphic transformations can be well approximated at the first layer as locally affine transformations or, in the limit, as local translations, which therefore falls into the partially observable groups case (see section 1.4). This assumption is substantiated by the following reasoning (here in 1-D for simplicity), in which any smooth transformation is seen as parameterized by the parameter r and expanded around zero as

$$T_r(I) = T_0(I) + A_r(I) + O(r) = (e + A_r)(I) + O(r),$$

where $e + A_r$ is a linear operator in $GL(\mathbb{R})$, the general linear group.

In our theory, the key property of hierarchical architectures of repeated HW modules—allowing the recursive use of single module properties at all layers—is the property of *covariance*.

It is illuminating to consider two extreme cartoon architectures for the first of the two stages described in figure 1.5:

- One layer comprising one HW module and its *KH* complex cells, each with a receptive field covering the whole visual field
- A hierarchy comprising several layers of HW modules with receptive fields of increasing size, followed by parallel modules, each devoted to invariances for a specific object class

In the first architecture, invariance to affine transformations is obtained by pooling over *KH* templates, each transformed in all possible ways: each of the associated simple cells corresponds to a transformation of a template. Invariance over affine transformation is obtained by pooling over the whole visual field. In this case, it is not obvious how to incorporate invariance to nongroup transformations directly into this one hidden layer architecture.

However, an HW module dealing with nongroup transformations can be added on top of the affine module. The theorems allow for this factorization (see appendix section A.9). Interestingly, they do not allow in general for factorization of translation and scaling (e.g., one layer computing translation invariance and the next computing scale invariance). Instead, the mathematics allows for factorization of the range of invariance for the same group of transformations. This justifies the first layers of the second architecture, corresponding to figure 1.5, stage 1, where the size of the receptive field of each HW module and the range of its invariance increases, from lower to higher layers.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

One of the main problems with the one-layer architecture is that it provides an invariant signature to an isolated object but not to its parts. This problem of recognizing wholes and parts is closely related to the problem of recognizing objects in clutter, recognizing an object independently of the presence of another one nearby. The key theorem about invariance assumes that image and templates portray isolated objects. Otherwise the signature may change because of different clutter at different times. Recognizing an eye in a face has the problem that the rest of the face is clutter. This is the old conundrum of recognizing a tree in a forest while still recognizing the forest.

A partial solution to this problem is a hierarchical architecture for stage 1 in which lower layers provide signatures with a small range of invariance for small parts of the image, and higher layers provide signatures with greater invariance for larger parts of the image (see appendix section A.3). All these signatures could then be used by class-specific modules, possibly in a reverse hierarchy strategy (see [42]), that is, using first the top-level signatures and then the low-level ones in a task-dependent, top-down mode. We describe this architecture, starting with the retina and V1, in chapter 3. Three points are of interest here:

- Factorization of the range of invariances is possible if a certain property of the hierarchical architecture, called covariance, holds. Assume a group transformation of the image that is, for instance, a translation or scaling of it. The first layer in a hierarchical architecture is called covariant if the pattern of neural activity at the output of the complex cells transforms according to the same group of transformations. It turns out that the architectures we describe have this property (see box 1.7): isotropic architectures, like the ones considered in this book, with pointwise nonlinearities, are covariant.
- Since each module in the architecture gives an invariant output if the transformed object is contained in the pooling range, and since the pooling range increases from one layer to the next, there is an invariance over larger and larger transformations. In order to make recognition possible for both parts and wholes of an image, the supervised classifier should receive signatures not only from the top layer (as in most neural architectures) but from the other levels as well (directly or indirectly).
- In the case of a discrete group one can prove that the number of different templates K required for selectivity is $K = C(\epsilon)(\log(n|G|)$ (see eq. (1.4)); that means it may significantly depend on the size of the group G that is pooled (see appendix section A.2.4).

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

1.8 Deep Networks and i-Theory

The class of learning algorithms called deep learning (particularly convolutional networks) is based on two computational operations in multiple layers.

The first operation is the inner product of an input with another point called a template (or filter or kernel), followed by a nonlinearity. The output of the dot product corresponds to the neural response of a simple cell [20]:

$$I \mapsto |\langle I, gt\rangle + b|_+,$$

where $|y|_+ = \max,(0, y)$.

The second operation is *pooling*. It aggregates in a single output the values of the different inner products computed, for example, via a sum

$$\sum_g |\langle I, gt\rangle + b|_+, \quad t \in \mathcal{T}, b \in \mathbb{R} \tag{1.9}$$

or a max operation $\max_g |\langle I, gt\rangle + b|_+$. This corresponds to the neural response of a complex cell [20]. In i-theory the nonlinearity acting on the dot product is considered part of the computation of a histogram, but the networks are completely equivalent.

It turns out (see references in [12]) that units of a deep convolutional network (DCN) using linear rectifiers (called by Breiman "ramps") correspond to a kernel with

$$\tilde{K}(I, I') = \int dg\, dg' \int dt\, db\, |\langle gt, I\rangle + b|_+ |\langle g't, I'\rangle + b|_+. \tag{1.10}$$

This result shows that linear rectifiers (and other nonlinear operations) in a deep network are equivalent to replacing the plain dot product with dot products in the feature space defined by the kernel. All the results described in the book hold for networks with a broad range of nonlinearities after the dot product.

Further extension are possible. Ongoing work [46] shows that i-theory can be formulated to include present-day deep convolutional learning networks (DCLNs) with supervision and nonpooling layers. In particular [46] shows how defining an extension of classical additive splines for multivariate function approximation (called hierarchical splines) is possible to have a theoretical framework for DCLNs with linear rectifers and pooling (sum or max).

The comparison of hierarchical versus shallow architecture in the context of i-theory and its extensions is explored in [47]. Hierarchical as well as shallow

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

networks can approximate functions of several variables, in particular those that are compositions of low-dimensional functions. [47] provides a characterization of the power of deep network architectures with respect to shallow networks: in particular it proves how shallow networks can approximate compositional functions with the same error of a deep network at the cost of a VC-dimension (Vapnik-Chervonkis dimension, a measure of the network capacity) that is exponential rather than quadratic in the dimensionality of the function. Compositional functions are also shown to be critical for image recognition, thus demonstrating a theoretical reason why deep architectures outperform shallow ones in image recognition.

Historical Background and Bibliography

There exists an extensive literature about invariant approaches to image representation. One of the first and most inspiring papers that used explicitly the group structure of the transformations and the group average technique to build invariant features is Mirbach's [43]. A different approach using group-invariant kernels was analyzed in Burkhardt [44]. Together with the work on kernel average embedding [45], it inspired our recent work on the equivalence between deep convolutional networks and hierarchical kernel machines [12]. More recently Mallat [32] developed an invariant representation called group scattering, which uses a cascade of modulus nonlinearities applied to wavelets coefficients in order to get group-invariant representations that are robust to small diffeomorphic transformations.

Chapter 1 has described work in our group. Its main sources are the following technical reports and journal papers: [1, 8, 9, 11, 12].

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

2 Biophysical Mechanisms of Invariance

Unsupervised Learning, Tuning, and Pooling

2.1 A Single-Cell Model of Simple and Complex Cells

At least two possible biophysical models for the Hubel-Wiesel module are implied by theory. The first is the original HW model of simple cells feeding into a complex cell. i-theory proposes the ideal computation of a cumulative distribution function (cdf), in which case the nonlinearity at the output of the simple cells is a threshold. A complex cell summing the outputs of a set of simple cells would then represent a bin of the histogram. A different complex cell in the same position pooling a set of similar simple cells with a different threshold would represent another bin of the histogram. Another possibility is that the nonlinearity at the output of the simple cells is approximately a square (or any power or combination of powers). In this case the complex cell pooling simple cells with the same nonlinearity would represent a moment of the distribution, including the linear average. Note that in this case some of the complex cells would be linear and would be classified by neurophysiologists using the standard criteria as simple. The nonlinear transformation at the output of the simple cells may correspond to the spiking mechanism in populations of cells (see references in [22]). The fact that a nonlinear biophysical mechanism can be approximated locally by a polynomial and thus can be used to compute combinations of moments has been noted several times [48, 49]. This is also the reason for the similar performance of models such as HMAX and the Freeman-Simoncelli model [48]. The nonlinearity in our model neurons (max or threshold) is well approximated by a few terms in a Taylor series approximation. As an aside, it should be emphasized that the distributions computed in our theory by complex cells do not have anything to do with the classical statistical interpretation of images and textures. The distributions do not come from the statistics of images but from the action of a group, and thus are independent of the nature of the images and whether or not they are textures.

The second biophysical model for the HW module that implements the computation required by i-theory consists of a single cell where dendritic branches play the role of simple cells (each branch containing a set of synapses with

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

weights providing, for instance, Gabor-like tuning of the dendritic branch) with inputs from the lateral geniculate nucleus (LGN). Active properties of the dendritic membrane distal to the soma provide separate thresholdlike nonlinearities for each branch separately, while the soma sums the contributions for all the branches. This model would solve the puzzle that so far there seems to be no morphological difference between pyramidal cells classified as simple versus complex by physiologists.

It is interesting that i-theory is robust with respect to the nonlinearity from simple to complex cells. We conjecture that almost any set of nontrivial nonlinearities will work. The argument rests on the fact that a set of different complex cells pooling from the same simple cells should compute the cumulative distribution or equivalently its moments or combinations of moments (each combination is a specific nonlinearity). Any nonlinearity will provide invariance if the nonlinearity does not change with time and is the same for all the simple cells pooled by the same complex cells. A sufficient number of different nonlinearities, each corresponding to a complex cell, can provide appropriate selectivity, assuming that each nonlinearity can be represented by a truncated power series and that the associated complex cells provide therefore a sufficient number of linearly independent combinations of moments.

2.2 Learning the Wiring in the Single-Cell Model

A simple model of how the wiring between a group of simple cells with the same tuning (e.g., representing the same eigenvector, having the same orientation) and a complex cell may develop is to invoke a Hebbian trace rule [49]. In a first phase, complex cells may have subunits with different selectivities (e.g., orientations), for instance, because natural images are rotation invariant and thus eigenvectors with different orientations are degenerate. In a second, plastic phase, subunits that are not active when the majority of subunits are active will be pruned out according to a Földiák-like rule [49]. Földiák's original trace rule says that the weight of a synapse between an input cell and an output cell is strengthened proportionately to the input activity and the trace or average of recent output activity at time t, where the dependence of the trace on previous activity decays over time with a certain decay constant.

2.3 Hebb Synapses and Principal Components

i-theory provides the following algorithm for learning the relevant invariances during unsupervised visual experience: storing sequences of images for each of a few objects (templates) while transforming (e.g., translating, rotating, and looming). We prove that in this way invariant hierarchical architectures

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

can be learned from unsupervised visual experience. Such architectures represent a significant extension, beyond simple translation invariance and beyond hardwired connectivity, of models of the ventral stream such as Fukushima's Neocognitron [2] and HMAX [47, 50] as well as deep neural networks called convolutional networks [34, 51] and related models [52–59].

In biological terms, the sequence of transformations of one template would correspond to a set of simple cells, each storing in its tuning a frame of the sequence. In a second learning step a complex cell would be wired to those simple cells. However, the idea of direct storage of sequences of images or image patches in the tuning of a set of V1 cells by exposure to a single object transformation is biologically rather implausible. Since Hebbian-like synapses are known to exist in the visual cortex, a biologically more plausible hypothesis is that synapses incrementally change over time as an effect of the visual inputs, that is, over many sequences of images resulting from transformations of objects (e.g., templates). The question is whether such a mechanism is compatible with i-theory, and how.

We explore this question for V1 in a simplified setup that can be extended to other areas. We assume the following:

1. The synapses between LGN inputs and (immature) simple cells are Hebbian, and in particular their dynamics follow Oja's flow. In this case, the synaptic weights will converge to the eigenvector with the largest eigenvalue of the covariance of the input images.
2. The position and size of the untuned simple cells are set during development according to an inverted pyramidal lattice (see figure 3.2 in the next chapter). The key point is that the size of the Gaussian spread of the synaptic inputs and the positions of the ensemble of simple cells are assumed to be set independently of visual experience.

In summary, we assume that the neural equivalent of the memorization of frames (of transforming objects) is performed online via Hebbian synapses that change as an effect of visual experience. Specifically, we assume that the distribution of signals seen by a maturing simple cell is Gaussian-like in x, y, reflecting the distribution on the dendritic tree of synapses from the lateral geniculate nucleus. We also assume that there is a range of Gaussian distributions with different σ, which increase with retinal eccentricity. As an effect of visual experience, the weights of the synapses are modified by a Hebb rule [60]. Hebb's original rule can be written as

$$\delta \mathbf{w}_n = \alpha y(\mathbf{x}_n)\mathbf{x}_n, \tag{2.1}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

where $\mathbf{x}_n$ is the input vector, $\mathbf{w}_n$ is the presynaptic weights vector, y is the postsynaptic response $y(\mathbf{x}_n) = y_n = \mathbf{w}_n^T \mathbf{x}_n$, and α is a positive constant called the learning rate. In order for this dynamical system to actually converge, the weights have to be normalized. In fact, there is considerable experimental evidence that the cortex employs normalization [61]. Hebb's rule, appropriately modified with a normalization factor, turns out to be an online algorithm to compute the eigenvectors of the covariance matrix of a set of input vectors. In this case it is called Oja's flow. Oja's rule [62, 63] defines the change in presynaptic weights $\mathbf{w}_n$ given the output response y of a neuron to its inputs to be

$$\Delta \mathbf{w}_n = \mathbf{w}_{n+1} - \mathbf{w}_n = \alpha y_n (\mathbf{x}_n - y_n \mathbf{w}_n). \tag{2.2}$$

The equation follows from expanding to the first-order Hebb rule normalized to avoid divergence of the weights.

Box 2.1
Principal Components

Principal component analysis (PCA) is a statistical procedure that allows identification of the principal directions in which the data vary. Those principal directions of variations are called principal components and are found by calculating the eigenvectors and eigenvalues of the data covariance matrix, defined as the matrix product between the data matrix and its transpose.

Since the Oja flow converges to the eigenvector of the covariance matrix of the $\mathbf{x}_n$ that has the largest eigenvalue, we are led to analyze the spectral properties of the inputs to simple cells and to study whether the eigenvectors (also called principal components (PCA); see box 2.1) can be used by the HW algorithm and in particular whether they satisfy the selectivity and invariance theorems.

Alternatives to Oja's rule can and should be considered [64, 65]. Note that a relatively small change in the Oja equation gives an online algorithm for computing independent components instead of PCs (see [66]). Which kind of plasticity is closer to biology remains an open question.

2.4 Spectral Theory and Pooling

Consider stage 1 of figure 1.5 in the previous chapter, which is retinotopic, and in particular the case of simple cells in V1. From assumption 1 in section 2.3, the lattice in space (x, y) and scale (s) of an immature simple cell is set during development of the organism (s is related to the variance of the Gaussian

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

envelope of the dendritic tree of the immature cell). Assume that all the simple cells are exposed while in a plastic state to a possibly large set of images $T = (t^1, \ldots, t^K)$. A specific cell at a certain position in x, y, s is exposed to the set of transformed templates $g_* T$, where g_* corresponds to the specific translation and scale that transforms the zero cell to the chosen neuron in the lattice. The associated covariance matrix is $C = g_* T T^T g_*^T$. Calling ϕ the first eigenvector of TT^T, we have that $g_* \phi$ is the first eigenvector of C. Thus it is possible to choose ϕ as a new template, and pooling over corresponding PCs across different cells (reached by different transformations from the zero cell) is equivalent to pooling over ϕ and its transformations. Both the invariance and selectivity theorems are valid. Empirically, we find [13] that the covariance of natural images yields eigenvectors that are Gabor-like wavelets with a random orientation for each size receptive field (second hypothesis in section 2.3). The random orientation follows from the argument together with the fact that the covariance of natural images is approximately rotation invariant. The Gabor-like shape can be qualitatively explained in terms of translation invariance of the correlation matrix associated with a set of natural images (and their approximate scale invariance, which corresponds to a $1/f$ spectrum; see also [67, 68]). Thus the Oja rule acting on natural images provides equivalent templates that are Gabor-like and are the optimal ones according to the theory of chapter 1 and appendix section A.2.5.

Consider now the nonretinotopic stage 2 of figure 1.5, in which transformations are not in scale or position, such as the transformation induced by a rotation of a face. Assume that a simple cell is exposed to all transformations g_i (g_i is a group element of the finite group G) of each set $T = (t_1, \cdots, t_K)$ of K templates. The cell is thus exposed to a set of images (columns of X) $X = (g_1 T, \cdots, g_{|G|} T)$. For the sake of this example, assume that G is a discrete group. Then the covariance matrix determining the Oja's flow is

$$C = XX^T = \sum_{i=1}^{|G|} g_i T T^T g_i^T. \tag{2.3}$$

It is evident that if ϕ is an eigenvector of C, then $g_i \phi$ is also an eigenvector with the same eigenvalue (for details on how receptive fields look in V1 and higher layers, see [11] and [69–71]). Consider, for example, G to be the discrete rotation group in the plane; then all the (discrete) rotations of an eigenvector are also eigenvectors. The Oja rule will converge to the eigenvectors with the top eigenvalue and thus to the subspace spanned by them. It can be shown that L^2 pooling over the PCs with the same eigenvalues represented by different simple cells is then equivalent to L^2 pooling over transformations (see appendix

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

section A.4). This argument can be formalized in the following variation of the pooling step in the HW algorithm.

2.4.1 Spectral Pooling Proposition

Suppose that Λ_k is the matrix corresponding to the group transformations of template t^k (each column is a transformation of the template). Consider the set of eigenvectors of the associated covariance matrix. Because of the preceding argument, $\langle g_m I, \phi_k \rangle = \langle I, \phi_p \rangle$, where $g_m^{-1}\phi_k = \phi_p$. In particular, it is easy to check that the first two moments can be computed either by pooling over eigenvectors or pooling over the transformations of the template (eq. (2.4)). The argument is still valid if the pooling is over part of the eigenvectors of $\Lambda = \cup_{i=1}^{K} \Lambda_i$. The first two moments, which can be thought of as part of an invariant signature, are

$$\mu_1 = \sum_i \langle I, \phi_i \rangle \,, \quad \mu_2(I) = \sum_i \langle I, \phi_i \rangle^2 \,; \tag{2.4}$$

to achieve invariance a complex cell can pool with a linear or quadratic nonlinearity over the eigenvectors of Λ_k instead of over the transformations of the template.

2.5 Tuning of Simple Cells

The theorems of chapter 1 on the HW module imply that the templates and therefore the tuning of the simple cells can be the image of any object. At higher levels in the hierarchy, the templates are neuroimages—patterns of neural activity—induced by actual images in the visual field. Section 2.4, however, offers a biologically more plausible way to learn the templates from unsupervised visual experience, via Hebbian plasticity. In chapter 3 we discuss predictions following from these assumptions for the tuning of neurons in the various areas of the ventral stream.

Background and Bibliography

This part of the book describes work in our group. Its main sources are the following technical reports: [9, 11].

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

3 Retinotopic Areas: V1, V2, V4

3.1 V1

3.1.1 A New Model from i-Theory for Eccentricity Dependence of Receptive Fields

The simplest and most common image transformations are affine transformations and in particular the subgroup consisting of shifts in position (x) and uniform changes in scale (s). The theory suggests that the first step of the invariance computation likely consists of learning the set of transformations. This is implicitly done via actual visual experience (or by evolution encoded in the genes or by a combination):

1. Store template t^k (which is an image patch and could be chosen at random).
2. Store all the observed transformed templates (bound together by continuity in time).
3. Repeat steps 1 and 2 for a set of K templates.

Although the templates can be arbitrary image patches, we assume—because of the optimal invariance theorem (see chapter 1) and because of the experimental evidence from V1 simple cells—that V1 templates are Gabor-like functions (more precisely, windowed Fourier transforms [72], p. 92).

Under scaling, a pattern exactly at the center of the fovea will change size without any translation of its center, and its boundaries will shift in x, y. For a pattern centered at some nonzero eccentricity, scaling will translate its center in the s, x plane (see figure 3.1). In the s, x plane the slope of the trajectory of a pattern under scaling is a straight line through the origin with a slope$s$ that depends on the size of the pattern and the associated position.

Consider a Gabor function at $s = s_0, x = 0$. s_0 is the minimum possible receptive field (RF) size, given optical constraints. Transforming it by shifts in x within $(-x_0, x_0)$ generates a set Λ_0 of templates. Suppose we want to ensure that what is recognizable at the highest resolution (s_0) remains recognizable at all scales up to $s_{\max}$.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

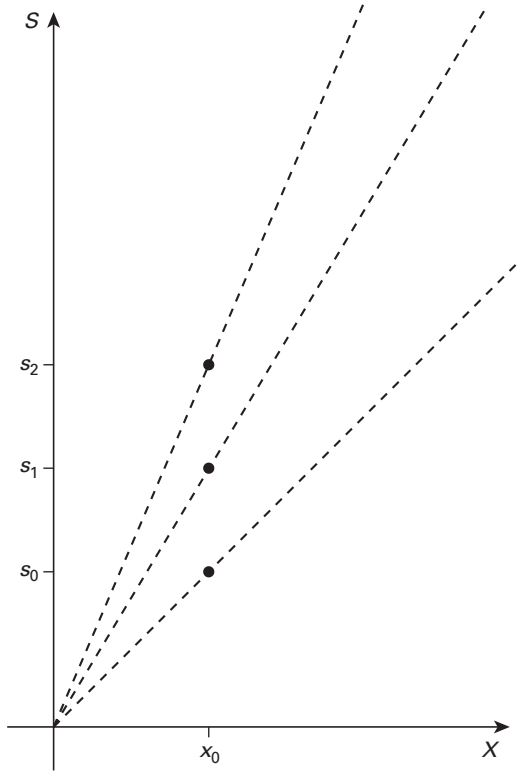

Figure 3.1

Left, the s, x plane, indicating templates of radius s_1, s_2, and s_3 in the spatial dimensions x, y, all centered at position X. While s and x are both measured in degrees of visual angle, in this plot the two axes are not shown to the same scale. Here, as in the rest of the book, we show only one spatial coordinate (x); everything we say can be directly extended to the x, y plane. We assume that the smallest template is the smallest simple cell in the fovea with a radius of around $40''$ [73]. For any fixed eccentricity (*right*) the size of the pattern determines the slope of its s, x trajectory under scaling. *Source:* [14].

The associated scale transformations of the set Λ_0 yield the inverted truncated pyramid shown in figure 3.2. Pooling over that set of transformed templates according to the Hubel-Wiesel algorithm (see chapter 1) will give uniform invariance to all scale transformations of a pattern over the range $(s_0, s_{\max})$; invariance to shifts will be at least within $(-x_0, x_0)$, depending on scale. The process of observing and storing a set of transformations of templates in order to be able to compute invariant representations may take place at the level of evolution or at the level of development of an individual system or as a combination of both.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

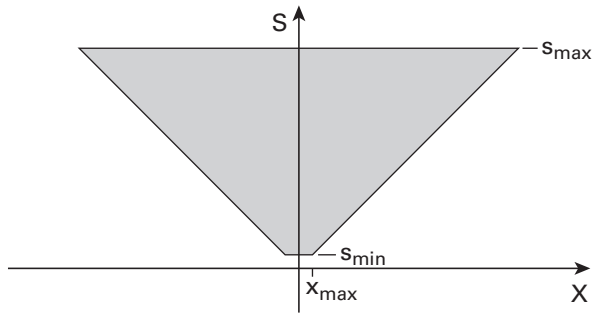

Figure 3.2

Synthesis of template set. Consider a template, such as a Gabor receptive field (RF), at $s_{\min}$ and $x = 0$. Store its transformations under (bounded) shift, filling the interval between $s_{\min}$, $x = 0$ and $s_{\min}$, $x_{\max}$. Store its transformations over (bounded) scale, filling the space in the inverted truncated pyramid. This is the space of bounded joint transformations in scale and space. For clarity, axes s, x are not in the same units (s unit is 10 times the x unit, so the real slope is $1/10$ of the slope shown). *Source:* [14].

The following definition holds. The inverted truncated pyramid of figure 3.2 is the locus Λ of the points such that their scaling between $s_{\min}$ and $s_{\max}$ gives points in S; further, all points P in $(-x_0, x_0)$ are in S, for example, $P \in \Lambda$ if

$$g_s P \in \Lambda, \quad s \in (s_{\min}, s_{\max}), \quad \Lambda_0 \in \Lambda, \tag{3.1}$$

where Λ_0 consists of all points at $s_{\min}$ between $-x_0$ and x_0. Other alternatives are possible: one is to set a constant difference between the minimum and the maximum scale at each eccentricity: $s_{\max} - s_{\min} = \text{const}$. Experimental data suggest this is a more likely possibility (large eccentricities are also represented in the visual system).

The new model consists of an inverted pyramid region. It follows naturally from i-theory if scale invariance has a higher priority than shift invariance. Recall that according to the optimal invariance theorem, the optimal template is a Gabor function. Under scale and translation within the inverted pyramid region, the Gabor template originates a set of Gabor wavelets (a tight frame). In the region of figure 3.2 scaling each wavelet generates other elements of the frame.

Note that the shape of the lower boundary of the inverted pyramid, although similar to standard empirical functions that have been used to describe the cortical magnification factor (M) scaling, is different for small eccentricities. An example of an empirical function [74, 75] is $M^{-1} = M_0(1 + ax)$, where M is the cortical magnification factor, M_0 and a are constants, and x is eccentricity.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A normal, nonfiltered image may activate all or part of the Gabor filters within the inverted truncated pyramid of figure 3.2. The pattern of activities is related to a multiresolution decomposition of an image. We call this transform of the image an inverted pyramid scale-space fragment (IP fragment) (the term *fragment* in this regard is borrowed from Ullman [76]) and consider it as supported on a domain in the space x, y, s that is contained in the inverted truncated pyramid of the figure. The fragment corresponding to a bandpass filtered image should be a more or less narrow horizontal slice in the s, x plane. In the following we assume that the template is a Gabor filter (of one orientation; other templates may have different orientations). We assume that the Gabor filter and its transforms under translation and scaling are roughly bandpass, and the sampling interval at one scale over x is s, implying half overlap of the filters in x. This is illustrated in figure 3.3. These assumptions imply that for each array of filters of size s, the first unit on the right of the central one is at $x = s$, if x and s are measured in the same units. For the scale axis we follow the sampling pattern estimated by Marr [73] with five frequency channels with $2s = 1'20'', 3.1', 6.2', 11.7', 21'$. Filter channels as described are supported by sampling by photoreceptors that starts in the center of the fovea at the Shannon rate, dictated by the diffraction-limited optics with a cutoff around 60 cycles/degree, and then decreases as a function of eccentricity.

3.1.2 Fovea and Foveola

In this truncated pyramid model of the simple cells in V1, the slope of the magnification factor M as a function of eccentricity depends on the size of the foveola, which we define here as the region at the minimum scale $s_{\min}$. The larger the foveola, the smaller the slope. We submit that this model nontrivially fits data about the size of the fovea, the slope of M, and the size of receptive fields in V1. In particular, the size of the foveola, the size of the largest RFs in the anterior in ferotemporal cortex (AIT), and the slope of acuity as a function of eccentricity depend on each other. Fixing one determines the other (after setting the range of spatial frequency channels, i.e., the range of RF sizes at $x = 0$ in V1). For a rough calculation we assume here that $s_{\min} \approx 40''$ (from an estimate of $1'20''$ for the diameter of the smallest simple cells [73, 77]). Data from Hubel and Wiesel [20, 78, figure A.6] and Gattass [79, 80], shown in figure 3.4, yield an estimate of the slope a for M in V1 (the slope of the lines $s_{\min}(x) = ax$). Hubel and Wiesel cite the slope of the average RF diameter in V1 as $a = 0.05$; Gattass quotes a slope of $a \sim 0.16$ (in both cases the combination of simple and complex cells may yield a biased estimate relative to the true slope of simple cells alone). Based on these actual data, our model of an inverted truncated pyramid

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

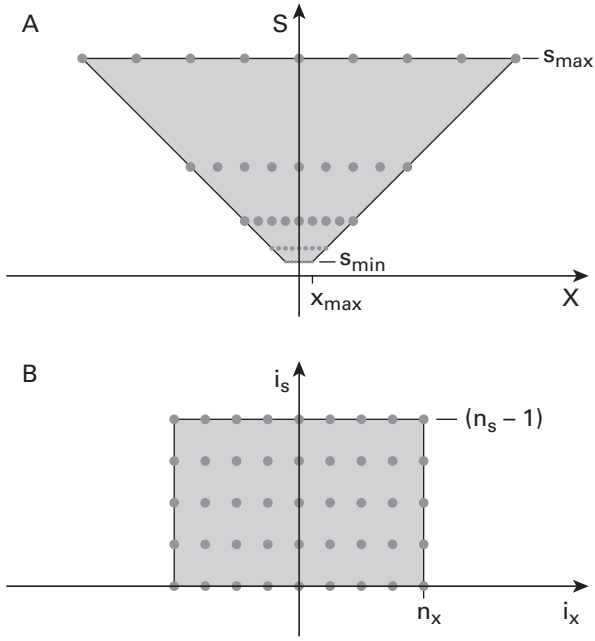

Figure 3.3

Under the assumption of Gabor filters and associated sampling for each scale at spatial intervals $\Delta x = s$, this figure depicts a subset of the resulting array of template units. The sampling over x follows the sampling theorem. There are no samples between the s axis and the line with slope 1 (when x and s are plotted in the same units). The center of the circles gives the s, x coordinates; the circles are icons symbolizing the receptive fields. The figure shows a foveola of $\approx 26'$ and a total of ≈ 40 units (20 on each side of the center of the fovea). It also shows sampling intervals at the coarsest scale in the fovea (assumed to be around $2s = 21'$ [73], which would span $\approx \pm 6$. The size of a letter on the ophthalmologist's screen for 20/20 vision is around $5'$. Since the inverted truncated pyramid (A) has the same number of sample points at every scale, it maps perfectly onto a square array (B) when x is replaced by $i_x = x/s$, the number of samples from the center. i_s is the scale band index (s, x units are scaled as in figure 3.2 for clarity). *Source:* [14].

predicts (using an estimate of $a = 0.1$) that the radius of the foveola (the bottom of the truncated pyramid) is $R = 1'20''/0.1 \approx 13'$ with a full extent of $2R \approx 26'$ corresponding to about 40 cells separated by $40''$ each. The size of the fovea (the top of the truncated pyramid) is predicted to then have $2R \approx 6'$ with 40 cells spaced $\approx 10'$ apart (see figure 3.3). Each of the scale layers has the same number of units, which is determined by the number of units in the foveola, that is, the number of units at the finest resolution. This remapping (see figure 3.3B) shows that $S1$ corresponds to a lattice of dimensions x, y, s, θ, where the dimension sizes are different (but roughly the same for x, y); the topology is that of a cylinder, with θ periodic.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

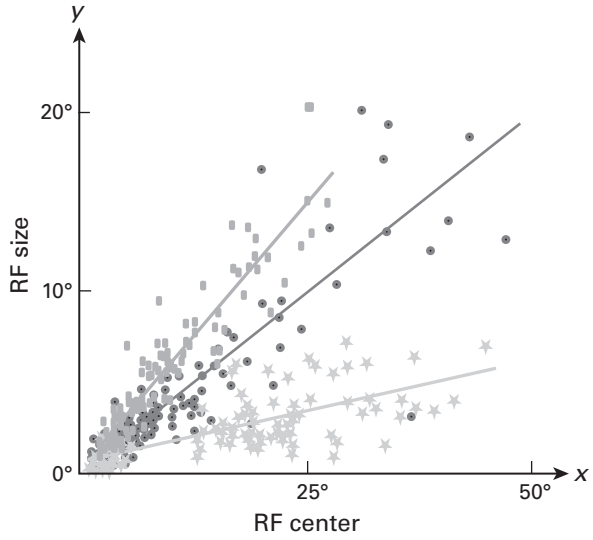

Figure 3.4

Data from Hubel and Wiesel [78] for monkey V1 give a slope for average RF diameter, relative to eccentricity, of $a = 0.05$. Data from other areas are similar but have higher slope (adapted from [48] with original monkey data from [79, 80]). *Source:* [14].

The lattice contains all the sampled translations and scale transformations of the templates required for invariance within the inverted pyramid. The number predicted by the theory is large but quite reasonable: $40 \times 40 = 1,600$ per scale and per Gabor orientation for a total of around 50K transformed templates.

To make these predictions, we used data from the macaque together with data from human psychophysics. Thus our estimates depend on the actual range of receptive field sizes and could easily be wrong by factors of 2. Our main goal is to provide a ballpark estimate of relevant quantities.

3.1.3 Scale and Position Invariance in V1

Invariance is not provided directly by the array but by the pooling over it. We limit ourselves in this section to the invariant recognition of isolated objects. The range of invariance in x is limited for each s by the slope of the lower bound of the inverted pyramid. The prediction is that the range of invariance Δx is proportional to the scale s, that is,

$$\Delta x \approx n_x s, \tag{3.2}$$

where n_x is the radius of the inverted pyramid and is the same for all scales s. Thus small details (high frequencies) have a limited invariance range in x,

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

whereas coarser details have larger invariance. n_x is obtained from the slope a of the cortical magnification factor as

$$n_x = \frac{1}{a}. \tag{3.3}$$

The following rationale for a so-called normative theory seems natural (and could easily be wrong):

1. Evolution tries to optimize recognition from few labeled examples,
2. As a consequence it has to optimize invariance,
3. As a first step it has to optimize invariance to scale and translation (under constraint of noninfinite resources),
4. Therefore it develops in V1 multiple sizes of receptive fields at each position, with RF size increasing linearly with eccentricity, properties that are reflected in the architecture of the retina.

3.1.4 Tuning of Cells in V1

Chapter 2 described the learning of templates through unsupervised experience of their transformations. The main example discussed consists of simple cells in V1. We can now add details to the example using the inverted pyramid architecture. Our basic hypothesis is that the position and size (a Gaussian distribution) of each immature simple cell is set by development and corresponds to a node in the lattice of figure 3.3 in x, y, s. Furthermore, Hebbian synapses on the simple cells will drive the tuning of the cell to be the eigenvector with the largest eigenvalue of the covariance of the input images. It can be shown ([11] and appendix section A.6) that because of the Gaussian envelope at each site in the lattice and because of the statistics of natural images, the tuning of each simple cell will converge to a Gabor-like function with properties that match very closely the experimental data on the monkey, the cat, and the mouse (see figure 1.4). Since the arguments in chapter 2 apply directly to this case, pooling over simple cells is equivalent to pooling over transformations of a template (a Gabor function with a specific orientation).

3.2 V2 and V4

3.2.1 Multistage Pooling

As discussed, pooling in one step over the whole s, x domain of (figure 3.3) suffers from interference from clutter-induced fragments in any location in the inverted pyramid. A better strategy is to pool area by area (in the ventral

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

stream), that is, layer by layer (in the corresponding hierarchical architecture). The C cells' activities in each layer of the hierarchy are sent to the memory or classification module at the top. After each pooling (C unit) stage, and possibly also after a dot product (S unit) stage, there is a downsampling of the array of units (in x, y and possibly in s) that follows from the low-pass-like effect of the operation. Note that according to i-theory, the templates in V2 and V4 should be patches of neural images at that level, possibly determined by PCA-like learning.

Here we analyze the properties of a specific hypothetical strategy of downsampling by 2 in x, y at each stage. We focus on shift invariance because the number of scales is so small (≤ 5) that pooling in scale could even be avoided all together. The choice of downsampling by 2 in each spatial dimension is simple and seems consistent with biological data. It is easy to modify the results by using the same logic with a different criterion for downsampling. Pooling each unit over itself and its neighbors (thus a patch of radius 1) allows downsampling of the array in x by a factor of 2. We assume that each combined S-C stage brings about a downsampling by 2 in each dimension of the x, y array. We call this process decimation; see figure 3.5.

Starting with V1, four stages of decimation reduce the number of units at each scale from ≈ 40 to ≈ 2, spanning $\approx 26'$ at the finest scale and $\approx 6'$ at the coarsest. Pooling over scale in a similar way may also decimate the array down to about one scale from V1 to IT. Neglecting orientations, in x, y, s the $30 \times 30 \times 6$ array of units may be reduced to just a few units in x, y and one in scale. This picture is consistent with the invariance found in IT cells [38]. According to i-theory, different types of such units are needed, one for each of several templates at the top level. Other types of pooling and downsampling are conceivable, such as pooling in space over each spatial frequency channel separately, possibly in different areas. Understanding the actual strategy underlying hierarchical pooling in the ventral stream is an open question. The simple strategy we describe is just a conjecture. Notice that downsampling in space by 2 at each stage, together with our previous estimates of the dimensions of the inverted pyramid in V1, predicts that about four to five layers in the hierarchy are required to obtain almost full invariance to shift and scale before stage 2. If layers are identified with visual areas, there should be at least four areas from V1 to AIT: it is tempting to identify them with V1, V2, V4, PIT.

3.2.2 Predictions of Crowding Properties in the Foveola and Outside It (Bouma's Law)

The pooling range is uniform across eccentricities in figure 3.6. The spatial pooling range depends on the area: for V1 it is the sampling interval between

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

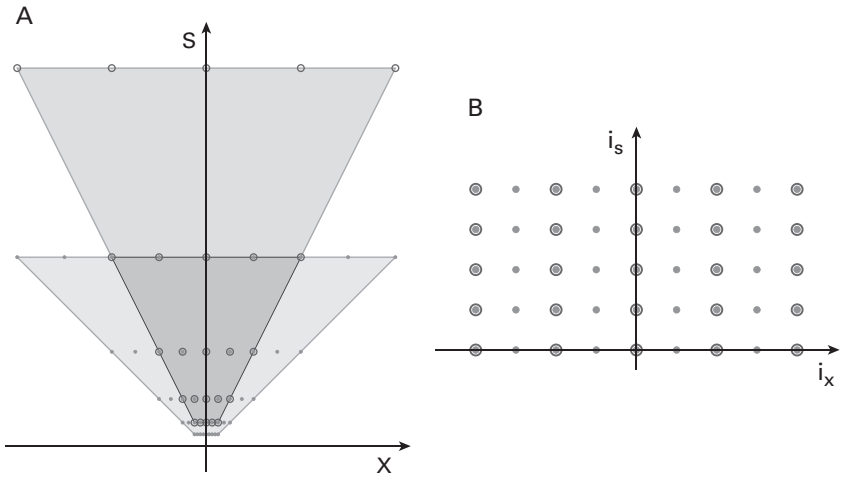

Figure 3.5

Pooling over 3×3 patches in x, y of the lattice of simple cells in V1 and subsampling decimates the lattice. If the lattice in x, y is 40 units, four steps (V1, V2, V4, PIT) of x pooling are sufficient to create cells that are 16 times larger than the largest in the fovea in V1 (probably around $21'$ at the coarsest scale), yielding cells with an RF diameter of up to ≈ 5. Each area in the fovea would see a doubling of size with corresponding doubling of the slopes at the border (before remapping to a cube lattice). The index of the units at position x and scale s is given by $i_x^s = 2^s i_x^1$. Simultaneous pooling over regions in S is possible. *Source:* [14].

the dots, for V2 it is the sampling interval between the shaded dots: they should be roughly equivalent to the radius of the receptive field of the complex cells in V1 and V2, respectively. V4 is not shown, but it is clear what is expected.

Figure 3.6 shows the regions of pooling corresponding to our assumptions. The inverted pyramid is split into sections, each one corresponding to pooling by a complex cell module. For now we consider only pooling in space and not scale. Figure 3.6 describes the situation for V1, but the same diagram can also be used for V2, V4, and PIT by taking into account the downsampling, the increase in size of the smallest cells, and the doubling in slope of the lower boundaries of the inverted pyramid. It is clear from the figure why the following criterion for pooling to remain interference-free from a flanking object, that is, unaffected by clutter, seems reasonable: the target and the flanking distractor must be separated by at least the pooling range and thus by a complex cell receptive field.

We consider two cases: (1) the target is in the central section at some layer (say, V2), and (2) the target is outside the foveola, that is, at an eccentricity greater than $10'$. Our predictions are as follows:

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

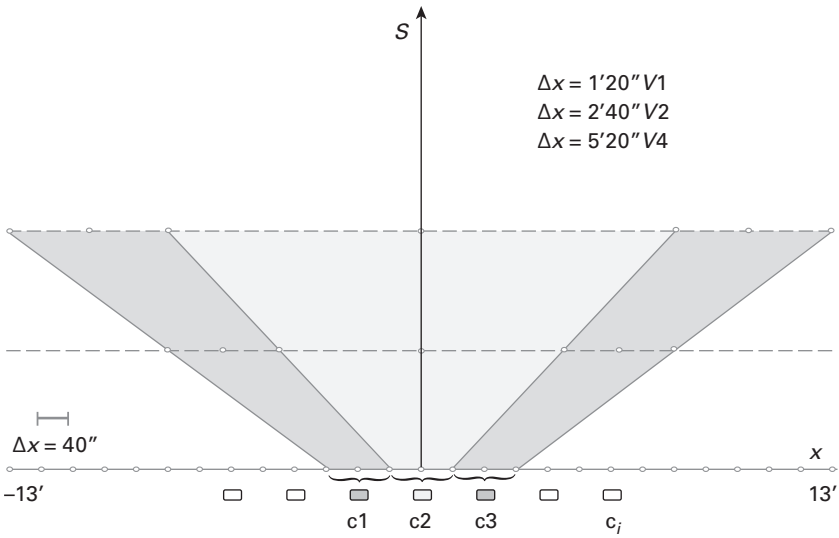

Figure 3.6

Part of the inverted pyramid of simple units in V1. The shaded regions are pooling regions, each one corresponding to a complex cell (there are more such regions to cover the extent of the foveola (diameter $\approx 21'$). The ones at the left and right boundaries will be at the edges of the inverted pyramid. Pooling shown here is spatial only (at each scale); it takes place around each simple unit and its two neighbors in x, y (a 3×3 patch). *Source:* [14].

- Consider a small target, such as a $5'$ width letter, placed in the center of the fovea, activating the smallest simple cells at the bottom of the inverted pyramid. The smallest critical distance to avoid interference should be the size of a complex cell at the smallest scale, that is, $\Delta x \approx 1'20''$ in V1 and $\Delta x \approx 2'40''$ in V2. Of course, in general, more than a scale will be activated. In particular, if the letter is made larger, the activation of the simple cells shifts to larger scales (s in figure 3.6), and since $\Delta x \approx s$, the critical spacing is proportional to the size of the target. It is remarkable that both these predictions match quite well figure 10 in Levi and Carney [81], although their data are for orientation discrimination.
- Outside the foveola, the size of the relevant complex cells determining pooling depends on spatial frequency content of target and distractors and on whether the pooling is only over space or also over scales (likely). Since the smallest RF size of the complex cells increases linearly with eccentricity, we expect that minimal critical spacing to avoid crowding outside to be roughly $\Delta x \approx bx$, with b depending on the cortical area representing the bottleneck. Thus the theory "predicts" Bouma's law [82] of crowding (often Bouma's

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

law is given as $\Delta x \approx bx + w$; see [83–85]). The experimental value found by Bouma for crowding of $b \approx 0.5$ points to a retinotopic area in the ventral stream, which may be V4, since the smallest slope found by Gattass for dependence on eccentricity of RF size is for V1 ≈ 0.12, for V2 ≈ 0.25, and for V4 ≈ 0.5 (on the other hand, studies of metameric stimuli by Freeman and Simoncelli [48] were interpreted by them to imply V2 in crowding and peripheral vision deficiencies).

3.2.3 Scale and Shift Invariance Dictates the Architecture of the Retina and the Retinotopic Cortex

It is interesting that from the perspective of i-theory, the linear increase of RF size with eccentricity, found in all primates, follows from the computational need of computing a scale- and position-invariant representation for novel images/objects. The theory also predicts the existence of a foveola and links its size to the slope of linear increase of receptive field size in V1. From this point of view, the eccentricity dependence of cone density in the retina as well as of the cortical magnification factor follow from the computational requirement of invariant recognition. The usual argument of limited resources (say, number of fibers in the optical nerve) does not determine the shape of the inverted pyramid but only the size of the fovea at the bottom (and thereby of the total number of cells). The inverted pyramid shape is independent of any bound on computational resources.

As mentioned, the estimate for the size of the foveola is quite small, with a diameter of $26'$ corresponding to about 40 simple cells (of each orientation) and about 50 cones in the retina. This is almost certainly a trade-off between limited translation invariance (the inverted pyramid region) and the ability to correct for it by moving gaze, while scale invariance is fully available in the center of the fovea. Notice that a fovea with a 10 times larger diameter would lead to an optical nerve of the same size as the eye, making it impossible to move it.

This state of affairs means that there is a quite limited field of vision in a single glimpse. Most of an image of 5×5 degrees is seen at the coarsest resolution only, if fixation is in its center. At the highest resolution only a very small fraction of the image (up to $30'$) can be recognized, and an even smaller part of it can be recognized in a position-invariant way (these numbers are rough estimates). An IP fragment can be defined as the information captured from a single fixation of an image. Such a fragment is supported on a domain in the space x, s, contained in the inverted truncated pyramid of figure 3.2. For normal-sized images and scenes with fixations well inside, the resulting

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

IP fragment will occupy most of the spatial and scale extent of the inverted pyramid.

Consider now the fragment corresponding to an object to be stored in memory (during learning) and recognized at runtime. Consider the most favorable situation for the learning stage: the object is close to the observer so that both coarse and fine scales are present in its fragment. At runtime the object can be recognized whenever it is closer or farther away (because of pooling). The important point is that a look at this and the other possible situations (at the learning stage) suggests that the matching should always weight more the finest available frequencies (bottom of the pyramid). This is the finding of Schyns and Gosselin [86]. As implied by their work, top-down effects may modulate these weights somewhat (this could be done during pooling), depending on the task. Assume that such a fragment is stored in memory for each fixation of a novel image. Then, because of large and position-independent scale invariance, there is the following trade-off between learning and runtime recognition:

- If a new object is learned from a single fixation, recognition may require multiple fixations at runtime to match the memory item (given its limited position invariance and unless fixation is set to be within the object).
- If a new object is learned from multiple fixations, with different fragments stored in memory each time, runtime recognition will need a lower number of fixations (in expectation).

The fragments of an image stored in memory via multiple fixations could be organized into an egocentric map. Though the map may not be directly used for recognition, it is probably needed to plan fixations during learning and especially during the recognition and verification stages (and thus indirectly used for recognition in the spirit of minimal images and related work by Ullman and coworkers,personal communication). Such a map may be related to Marr's $2\frac{1}{2} = D$ sketch [87], for which no neural correlate has been found as yet.

Thus, simultaneous invariance to translation and scale leads to an inverted truncated pyramid as a model for the scale-space distribution of the receptive fields in V1. This is a new alternative, as far as we know, to the usual smooth empirical fit of the cortical magnification factor data. In particular, the linear dependence on eccentricity of receptive field sizes in V1 (and cones sampling in the retina) follows from the computational requirement of scale invariance. This picture contains other interesting properties: the range of shift invariance depends on scale; the size of the flat, high acuity foveal region, which we identify with the foveola, can be inferred from the slope of the eccentricity-dependent acuity. Existing neural data from macaque V1 suggest a foveola with a diameter of around $20'$ of arc. In a sense, scale invariance turns out to

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

be more natural than shift invariance in this model (there is scale invariance over the full range at the fixation point). This is to be expected in organisms in which eye fixations can easily take care of shift transformations, whereas more extensive motions of the whole body would be required to change the scale. i-theory predicts scale invariance at this level: this is exactly what Anstis [87] found: a letter just recognizable at some eccentricity remains equally recognizable under scaling.

Physiology data suggest that AIT neurons are somewhat less tolerant to position changes of small stimuli [88, 89]. A comparison across studies suggests that position tolerance is roughly proportional to stimulus size [89]. If we assume that some IT neurons effectively pool over all positions and scales, i-theory in fact expects that their invariant receptive field (over which consistent ranking of stimuli is maintained) should be smaller for higher spatial frequency patterns than for low-frequency ones. There is evidence of attentional suppression of nonattended visual areas in V4. From the perspective of i-theory, it seems natural that top-down signals may be able to control the extent of pooling, or the pooling stage, used in computing a signature from a region of the visual field, in order to minimize clutter interference.

In summary, i-theory provides explanations and computational justifications for several known properties of the retinotopic cortex. It also makes a few predictions that are still waiting for experimental tests:

- There is an inverted pyramid of simple cells size and positions with parameters specified in the text, including linear slope of the lower boundaries. The predicted pyramid is consistent with available data. More precise measurements in the region of the foveola could decide between the usual empirical fits and our predictions.
- Anstis [87] did not take measurements close to the minimum letter size, which is about $5'$ for $20/20$ vision. i-theory predicts that if there is a range of receptive fields in V1 between $s_{\min}$ and $s_{\max}$ in the fovea, then there is a finite range of scaling between $s_{\min}$ and $s_{\max}$ under which recognition is maintained [14]. It is clear that looking at the image from an increasing distance will at some point make it unrecognizable; it is somewhat less clear that getting too close will also make it unrecognizable (this phenomenon was found in Ullman's minimal images [Ullman, personal communication]).
- Consider the experimental use of images such as novel letters (never seen before) of appropriate sizes that are bandpass filtered (with the Gabor-like filters assumed for V1). The prediction—because of the pooling over the whole inverted pyramid (done between V1, V2, and V4) — is that for a new presentation there will be psychophysical scale invariance for all frequencies

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

between $s_{\min}$ and $s_{\max}$. Shift invariance increases linearly with spatial wavelength and is at any spatial frequency at least between $x_{\min}$ and $x_{\max}$ (the bottom edge of the truncated pyramid).

- i-theory predicts a flat region of constant maximum resolution, which we called the foveola. Its size determines the slope of the lower border of the pyramid. Since the slope can be estimated relatively easily from existing data, our prediction for the linear size of the foveola is about $40'$ of arc, corresponding to about 30 simple cells of the smallest size (assumed to be $\approx 1'20''$ of arc). Note that our definition of the fovea is in terms of the set of all scaled versions of the foveola between $s_{\min}$ and $s_{\max}$ spanning about 6 degrees of visual angle.
- i-theory explains crowding effects in terms of clutter interference in the pooling stage (see also [90]). It predicts Bouma's law and its linear dependence on eccentricity [82]. Since Bouma's constant has a value of about 0.4 (see [48]), our theory requires that a signature that is interference-free from clutter at the level of V2 is critical for recognition. This is consistent with the i-theory independent requirement [8, 9] that signals associated with image patches of increasing size from different visual areas must be able to access memory and classification stages separately. The requirement follows from the need to recognize parts and wholes in an image and to avoid clutter for small objects. The V2 signal could directly or indirectly (via IT, or V4 and IT) reach memory and classification.
- The angular size of the fovea remains the same at all stages of a hierarchical architecture (V1, V2, V4,...), but the number of units per unit of visual angle decreases and the slope increases because the associated x increases (see figure 3.5). The theory predicts crowding in the foveola (very close to fixation) but with very small Δx (see eq. (3.2)) that depends on the size of the (small) objects rather than eccentricity. For objects smaller than $20'$ in diameter, the prediction (to be tested) is $\Delta x \approx 3' - 4'$, assuming that the main effect of clutter is in V2 and at the smallest of the five channels of simple cells.

3.2.4 Tuning of Simple Cells in V2 and V4

It is difficult to make clear predictions about V2 and V4 without additional specific data because several options are allowed by the theory. A simple scenario is as follows. There is an x, y, s lattice for V2 (and another for V3) simple cells, as shown in figure 3.5. The tuning of each simple cell (a point in the x, y, s lattice) is determined by the top principal component computed on the neural activity of the complex cells in V1, seen through a Gaussian window that

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

includes $\approx 3 \times 3$ complex cells in x, y and a set of orientations for each position and scale. It is likely—and supported by preliminary experiments [11]—that some of the principal components computed in this way over a large number of natural images can lead to cell tunings similar to measurements in V4.

Background and Bibliography

This part of the book describes work in our group. Its main source is the technical report [14].

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

4 Class-Specific Approximate Invariance in Inferior Temporal Cortex

4.1 From Generic Templates to Class-Specific Tuning

As discussed in chapter 1, approximate invariance for transformations beyond the affine group requires highly tuned templates and therefore highly tuned simple cells. This is expected to take place in higher, nonretinotopic visual areas of the hierarchy. This is consistent with the architecture of the ventral stream and the the existence of class-specific modules in the primate cortex such as a face module and a body module [91–94]. We saw that areas in the hierarchy up to V4 or PIT provide signatures for larger parts or full objects. Thus we expect the following:

- Inputs to the class-specific modules are more scale and shift invariant than in earlier areas such as V1.
- Class-specific templates are large. For instance, in the case of faces, templates should cover significant regions of the face. Note that only large templates support pose invariance: the image of an isolated eye does not change much under rotations in depth of the face (and at most it undergoes scaling and shift).

4.2 Development of Class-Specific and Object-Specific Modules

A conjecture emerging from i-theory offers an interesting perspective on anterior inferior temporal cortex (AIT) [13]. For transformations that are not affine transformations in 2-D (we assume 3-D information is not available to the visual system or used by it, which may not always be true), an invariant representation cannot be computed from a single view of a novel object because the information available is not sufficient: lacking are the 3-D structure and the material properties of the object. Thus exact invariance to rotations

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

in depth or to changes in the direction or spectrum of the illuminant cannot be obtained. However, as shown in chapter 1, approximate invariance to smooth nongroup transformations can still be achieved in several cases (but not always) using a standard Hubel-Wiesel (HW) module. The reason this will often approximately work is because it effectively exploits prior knowledge of how similar objects transform. The image-to-image transformations caused by a rotation in depth are not the same for two objects with different 3-D structures. However, objects that belong to an object class where all the objects have similar 3-D structures transform their 2-D appearance in (approximately) the same way. This commonality is exploited by an HW module to transfer the invariance learned from (implicitily supervised) experience with template objects to novel objects seen only from a single example view. This is effectively our definition of an object class: *a class of objects such that the transformation for a specific object can be approximately inferred from how other objects in the class transform*. The necessary condition for this to hold is that the 3-D shapes be similar between any two objects in the class. The simulation in figure 4.1 shows that HW modules tuned to templates from the same class of (always novel) test objects provide a signature that tolerates substantial viewpoint changes (plots on the diagonal). It also shows the deleterious effect of using templates from the wrong class (plots off the diagonal). There are, of course, several other class-specific transformations besides rotation in depth, such as face expression and body pose transformations, to which the same arguments apply.

An interesting conjecture is that the visual system is continuously and automatically clustering objects and their transformations, observed in an unsupervised or implicitely supervised way, into class-specific modules [13]. Images of an object and of its transformations correspond to a new orbit Λ_k. New images are added to an existing module only if their transformations are well predicted by it. If no module can be found with this property, the new orbit will be the seed of a new object cluster/module.

For the special case of rotation in depth, Leibo et al. [13] ran a simulation using 3-D modeling, rendering software to obtain the orbits of objects for which there exist 3-D models. Faces had the highest degree of clustering of any natural category, unsurprising since recognizability likely influenced face evolution. A set of chair objects had broad clustering, implying that little invariance would be obtained from a chair-specific region. A set of synthetic wire objects, very similar to the paperclip objects used in several classic experiments on view-based recognition (e.g., [95–97]) were found to have the smallest index of compatibility [13]. Experience with familiar wire objects does not

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

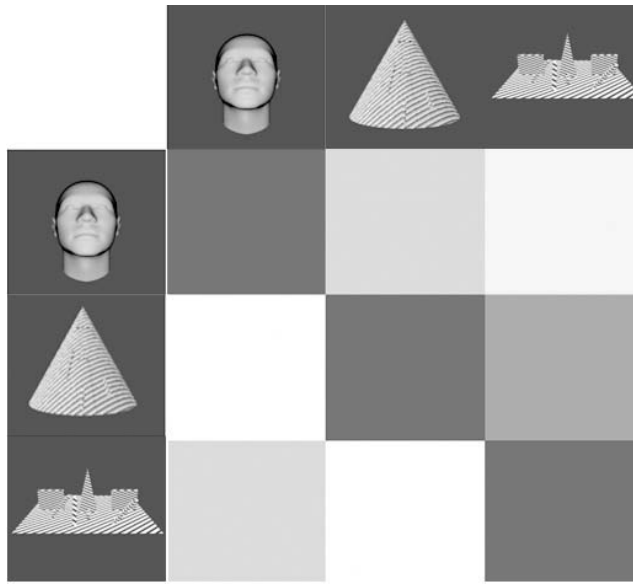

Figure 4.1

Class-specific transfer of depth rotation invariance for images from three classes: faces (A), cylinders (B), and composed (C). The left column of the matrix shows the results of the test for invariance for a random image of a face (class A) in different poses with respect to 3-D rotation using 3-D rotated templates from (classes A, B, and C). Similarly, the middle and the right columns show the invariance results for classes B and C, tested on rotated templates of A, B, and C, respectively. The different gray shades in the matrix show the maximum invariance range (degrees of rotation away from the frontal view). Only the diagonal values of the matrix (train A–test A; train B–test B; train C–test C) show an improvement of the view-based model over the pixel representation. That is, only when the test images transform similarly to the templates is there any benefit from pooling [13].

transfer to new wire objects because the 3-D structure is different for each individual paperclip object.

It is instructive to consider the limit case of object classes that consist of single objects like individual paperclips. If the object is observed under rotation, several frames are memorized as transformations of a single template (identity is implicitly assumed to be conserved by a Földiák-like rule, as long as there is continuity in time of the transformation; this is a case of implicitly supervised learning). The usual HW module pooling over them will allow view-independent recognition of the specific object. A few comments:

- Remarkably, the HW module for class-specific transformations, when restricted to multiple-views, single-object, is equivalent to the Poggio-Edelman model [98] for view invariance.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

- The class-specific module is also effectively a gate. In addition to providing a degree of invariance it performs a template-matching operation with templates that can effectively block images of other object classes. This gating effect may be important for the system of face patches discovered by Freiwald and Tsao [105, 106], and it is especially obvious in the case of a single-object module.
- From the point of view of evolution, the use of the HW module for class-specific invariance can be seen as a natural extension from its role in single-object view invariance. The latter case is computationally less interesting, since it effectively implements a look-up table, albeit with interpolation power. The earlier case is more interesting, since it allows generalization from a single view of a novel object. It also represents a clear case of *transfer of learning*.
- Some of the best evidence for modules in IT for individual objects is provided by the Logothetis data from monkeys who were trained intensively to specific unfamiliar paperclip objects [96–98]. Unfortunately, fMRI data about spatial clustering in cortex was not available at the time.

4.3 Domain-Specific Regions in the Ventral Stream

There are other domain-specific regions in the ventral stream besides faces and bodies [13]. We think that additional regions for less common or less transformation-compatible object classes will appear with higher resolution imaging techniques. One example may be the fruit area, discovered in macaques with high-field fMRI [99]. Others include the body area and the lateral occipital complex (LOC), which is not really a dedicated region for general object processing but a heterogeneous area of cortex containing many domain-specific regions too small to be detected with the resolution of fMRI [100]. The visual word form area (VWFA) [101] seems to represent printed words. In addition to the generic transformations that apply to all objects, printed words undergo several nongeneric transformations that never occur with other objects. For instance, reading is rather invariant to font transformations and can deal with handwritten text. Thus, VWFA is well accounted for by the invariance hypothesis. Words are frequently viewed stimuli that undergo class-specific transformations. i-theory suggests that the limit case is object classes that consist of single objects. We expect that IT will include many such small modules in addition to large ones such as the face and the body modules. This prediction is consistent with evidence (e.g., [102]) of clustering of tuning in IT by category. Our prediction does not depend on supervised learning.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

The justification—actually a prediction—by i-theory for domain-specific regions in cortex is different from other proposals. However, it is in complementary with respect to some of them, rather than exclusive. For instance, it would make sense that the clustering depends not only on the index of compatibility but also on the relative frequency of each object class. The conjecture claims that transformation compatibility is the critical factor driving the development of domain-specific regions and that there are separate modules for object classes that transform differently from one another.

4.4 Tuning in the Inferior Temporal Cortex

Considering the limit case of object classes that consist of single objects is important in order to understand the functional architecture of most of IT. If a specific object is observed under a set of transformations, several images of it can be memorized and linked together by continuity at the time of the transformation. As we mentioned, the usual HW module pooling over them will allow view-independent recognition of the specific object. Since this is equivalent to the Poggio-Edelman model for view invariance [98], there is physiological support for this proposal (see [97, 103, 104]).

4.5 Mirror-Symmetric Tuning in the Face Patches and Pooling over Principal Components

The theory can explain wether its framework the Freiwald-Tsao data [105, 106] on the face patch system. The most posterior patches, middle lateral and fundus (ML/MF), provide view- and identity-specific input to the anterior lateral patch (AL), where most neurons show tuning that is an even function of the rotation angle around the vertical axis. The anterior medial patch (AM), which receives inputs from AL, is identity specific and view invariant. The puzzling aspect of these data is the mirror-symmetric tuning in AL. Why does this appear in the course of a computation that leads to view invariance?

The theoretical framework implies that neurons in patch AL [105, 107], when exposed to several different faces, each one generating several images corresponding to different rotations in depth, may become tuned to the eigenvectors corresponding to the set of views. Since faces are bilaterally symmetric and since for each view there is a mirror-symmetric view, the associated covariance function has eigenvectors (PCs) that are either even or odd functions (see figure 4.2 and appendix section A.5).

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

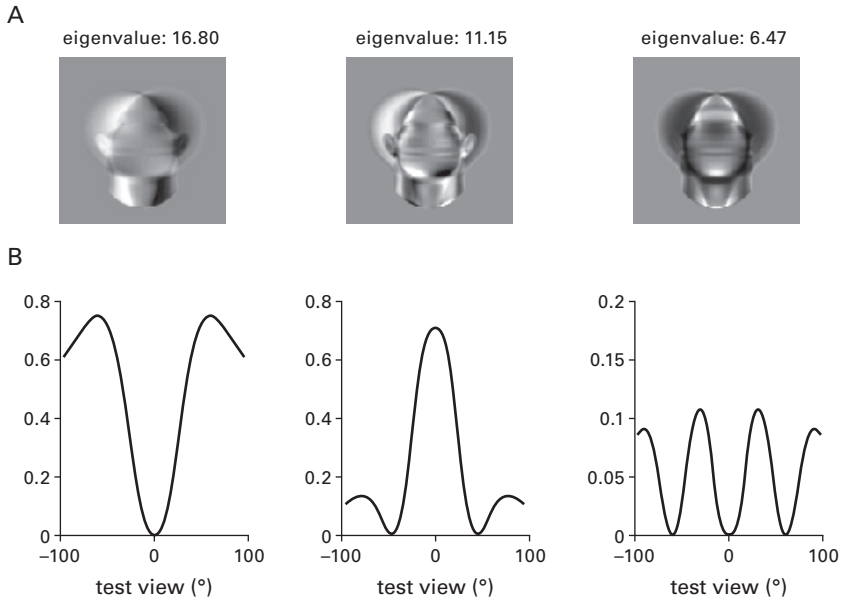

Figure 4.2

Face identity is represented in the macaque face patches [105]. Neurons in the middle areas of the ventral stream face patch (ML, MF) are view specific, while those in the most anterior area (AM) are view invariant. Neurons in an intermediate area (AL) respond similarly to mirror-symmetric views. In i-theory view invariance is obtained by pooling over simple neurons whose tuning corresponds to the PCA of a set of faces previously experienced each under a range of poses. Because of the bilateral symmetry of faces, the eigenvectors of the associated covariance matrix are even or odd. This is shown in (A), where the first three PCs of set of grey-level faces under different poses are plotted. The same symmetry arguments apply to neural images of faces. (B) Response of three model AL units to a face stimulus as a function of pose under different poses [13].

The observed AL properties could then be explained in different ways.

- The class-specific theorem and the spectral pooling proposition (see chapter 2) suggest that square pooling (i.e., energy pooling) over these face PCs can provide approximate invariance to rotations in depth. The full argument goes as follows. Rotations in depth of a face around a certain viewpoint, say, $\theta = \theta^0$, can be approximated well by linear transformations (by transformations $g \in GL(2)$). The HW algorithm can then provide invariance around $\theta = \theta^0$. Finally, if different sets of simple cells are plastic at somewhat different times, exposure to a partly different set of faces yields different eigenvectors summarizing different sets of faces. The different sets of faces play the role of different object templates in the standard theory. According to the theory, the result should be expected if AL contains

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

simple cells that are tuned by a synaptic Hebb-like Oja rule and the output of the cells is roughly a squaring nonlinearity as required by the spectral pooling proposition. In this interpretation, cells in AM pool over several of the squared eigenvector filters to obtain invariant second moments (see figure 4.2). Detailed models from V1 to AM show properties that are consistent with the data. It is doubtful, however, that such models can have a sufficiently good performance in face recognition [7, 15, 35, 108].

- An alternative explanation is that the network leading to each AL cell is balanced [108] and the cell has a mirror-symmetric tuning (induced by a set of template faces averaged over transformations).

Background and Bibliography

This part of the book describes work in our group. Its main sources are the following technical reports: [9, 11, 13, 15].

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

5 Discussion

5.1 i-Theory: Main Ideas

Neurally plausible computation for learning invariant representations. i-theory deals with a key computational problem in object recognition: identifying or categorizing an object after looking at a single example of it or of an exemplar of its class. To paraphrase Geman [41], the difficulty in understanding how biological organisms learn—in this case how they recognize—is not the usual $n \to \infty$ but $n \to 0$. The invariance algorithm we study is suggested by known properties of neurons, in particular their capability of performing almost automatically an approximation of the dot product of a vector of inputs with another stored vector (of synaptic weights). We introduce the notion of a *signature* of an object as a set of similarity measurements with respect to the group transformations of a small set of template images. We prove that the signature vector is invariant for the same group of transformations. This result provides a solution for the problem of recognizing an object after training with a single image of it—under conditions that are idealized but hopefully capture a good approximation of reality.

The study of this algorithm and its neural implementation leads to a number of predictions and explanations related to aspects of the ventral stream. The main ones are as follows:

- A two-stage architecture comprising a retinotopic set of areas with cell tuning that is independent of the object class and a set of nonretinotopic, class-specific modules
- Increasing receptive field (RF) sizes and invariance in the areas of the retinotopic stage
- Object-specific invariances in the class-specific modules

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

- Gabor-like RFs in V1 because of optimal simultaneous invariance in scale and position
- A new model of eccentricity dependence of RFs in cortical areas because of the need for position and scale invariance

5.2 i-Theory, Deep Learning Networks, and the Visual Cortex

Since the work on i-theory began, a class of learning algorithms called deep convolutional learning networks (DCLNs)—a new name for multilayer perceptrons (MLPs)—has achieved good performance in many applications. Historically, hardwired invariance to translation was first introduced in the Neocognitron by Fukushima [2] and later in LeNet [109] and in HMAX [47]; HMAX also had invariance to scale. These architectures are early examples of convolutional networks. There is no significant difference between i-theory and DCLNs. The basic operation in both cases is obtaining the inner product of the input I with a group transformation gt of a template. A pooling stage follows. Thus the basic operation in a convolutional layer is

$$\sum_g \sigma(\langle I, gt\rangle + b), \quad t \in \mathcal{T}, b \in \mathbb{R}. \tag{5.1}$$

If σ is a step function, the pooling can compute a histogram. However, several other nonlinearities are admissible, being invariant and selective (see chapter 1 and [9]). The nonlinearity used in present-day DCLNs is a "ramp," that is, a rectification nonlinearity.

$$\sum_g |\langle I, gt\rangle + b|_+, \quad t \in \mathcal{T}, b \in \mathbb{R}. \tag{5.2}$$

Note that nonconvolutional layers are a special case of the preceding equations when the only element in G is the identity. The rectification is neurally quite plausible, and it should be included in a theory of the visual cortex. As mentioned, the results of i-theory continue to hold:

- Linear combinations of rectified dot products are equivalent to kernels measuring the similarity between I and gt. Thus the effect of rectifying nonlinearities is to define a new dot product in the feature space of the kernel.
- Group averages of nonlinear functions of the dot products are invariant and selective (no need for explicit histograms).

The *pooling* operation

$$\sum_g |\langle I, gt\rangle + b|_+, \quad t \in \mathcal{T}, b \in \mathbb{R}. \tag{5.3}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

can also take other forms such as

$$\max_g |\langle I, gt\rangle + b|_+. \tag{5.4}$$

An alternative form of pooling, similar but biologically more plausible, is provided by softmax pooling:

$$\sum_g \frac{(\langle I, gt\rangle)^n}{\sum_{g'}(1 + \langle I, g't\rangle)^{n-1}}, \quad t \in \mathcal{T}, \tag{5.5}$$

Present-day DCNs seem to use eq. (5.5) for the pooling layers and eq. (5.2) for the nonconvolutional layers. The latter is the degenerate case of eq. (5.5) (when G contains only the identity element).

5.3 Predictions and Explanations

From the point of view of neuroscience, i-theory provides a number of predictions and explanations. Following are questions that the theory tries to answer:

- What is visual cortex computing?
- What are the functions and circuits of simple and complex cells?
- Why is there Gabor-like tuning in simple cells?
- Why is there generic Gabor-like tuning in early areas and specific selective tuning higher up?
- What is the computational reason for the eccentricity-dependent size of RFs in V1, V2, V4?

Some predictions are the following:

- Simple and complex cells should be found in all visual and auditory areas, not only in V1. Our definitions of simple cells and complex cells are different from the traditional ones used by physiologists. In i-theory complex cells represent invariant measurements associated with statistics of the outputs of simple cells. The theory implies that under some conditions exact or approximate invariance to all geometric image transformations can be learned, either during development or in adult life. It is, however, also consistent with the possibility that basic invariances may be genetically encoded by evolution but refined and maintained by unsupervised visual experience. A single-cell model for simple and complex cells follows from the theory as an interesting possibility.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

- The output of V2, V4, PIT should be capable of accessing memory either via connections that bypass higher areas or indirectly via equivalent neurons in higher areas (because of the argument about clutter).
- Areas V1, V2, V4, and possibly PIT are mainly dedicated to compute signatures that are invariant to translation, scale, and their combinations as experienced in past visual experience.
- Inferotemporal cortex is a complex of parallel class-specifc modules for a large number of large and small object classes. These modules receive position- and scale-invariant inputs (invariance in the inputs greatly facilitates unsupervised learning of class-specific transformations). From the perspective of the theory, the logothetis data [97] concern single-object modules and strongly support the prediction that exposure to a transformation leads to neuronal tuning to several frames of it.
- Each cell's tuning properties are shaped by visual experience of image transformations during developmental and adult plasticity.
- The mix of transformations (seen from the retina) learned in each area influences the tuning properties of the cells: oriented bars in V1 and V2, radial and spiral patterns in V4, class-specific tuning in anterior IT (e.g., face-tuned cells).
- During evolution areas above V1 should appear later than V1, reflecting increasing object categorization abilities and the need for invariance beyond translation.
- An architecture based on signatures that are invariant (from an area at some level) to affine transformations may underlie *perceptual constancy* against small eye movements and other small motions. There may be physiological evidence [110] suggesting invariance of several minutes of arc at the level of V1 and above.
- Invariance to affine transformations (and others) can provide the seed for evolutionary development of conceptual invariances.
- The *transfer of invariance* accomplished by the machinery of the set of transformed templates may be used to implement high-level abstractions.

5.4 Remarks

Object-Based versus 3-D versus View-Based Recognition

We mention here an old controversy about whether visual recognition is based on views or on 3-D primitive shapes called geons. According to i-theory, image views retain the main role, but ideas related to 3-D shape may also be valid. The psychophysical experiments of Edelman and Bülthoff [29, 96] concluded

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

that generalization for rotations in depth was limited to a few degrees ($\approx \pm 30$ degrees) around a view independently of whether 2-D or 3-D information was provided to the human observer (psychophysics in monkeys [96, 97] yielded similar results). The experiments were carried out using paperclip objects with random 3-D structures (or similar but smoother object). For this type of objects class-specific learning is impossible (they do not satisfy the second condition in the class-specific theorem), and thus i-theory predicts the result obtained by Edelman and Bülthoff. For other objects, however, such as faces, the generalization that can be achieved from a single view by i-theory can span a much larger range than ± 30 degrees, indirectly exploiting 3-D-like information from templates of the same class.

Genes or Learning

i-Theory shows how the tuning of the simple cells in V1 and other areas could be learned in an unsupervised way. It is possible, however, that the tuning (or better, the ability to quickly develop it in interaction with the environment) may have been partly compiled during evolution into the genes. Note that this hypothesis implies that most of the time the specific function is not fully encoded in the genes; genes facilitate learning but do not replace it completely. It has to be expected in the nature versus nurture debate that nature usually needs nurture, and nurture is made easier by nature.

Computational Structure of the Hubel-Wiesel Module

The HW module computes the cumulative distribution function of $\langle I, g_i t^k \rangle$ over all $g_i \in G$ or equivalent statistics. The computation consists of

$$\mu_h^k(I) = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(\langle I, g_i t^k \rangle + h\Delta). \tag{5.6}$$

The main forms of the nonlinearity σ are either a threshold function or a power $n = 1, \ldots, \infty$ of its argument. Several known networks are special cases of this module. One interesting case is when G is the translation group and $\sigma(\cdot) = \|\cdot\|_\infty$; then the equation is related (for $H = 0$) to a unit in a convolutional network with max pooling. In another noteworthy case (we always assume that I and t^k are normalized) the equation is very similar to the radial basis function network proposed by Poggio and Edelman [98] for view classification. In this spirit, note that the equation for a unit in a convolutional network is

$$\frac{1}{|G|} \sum_{i=1}^{|G|} c_i \sigma(\langle I, g_i t \rangle + h\Delta), \tag{5.7}$$

where I is the input vector, c_i, t, Δ are parameters to be learned in supervised mode from labeled data, and $g_i t(x) = t(x - i\delta_x)$. Thus units in convolutional

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

networks are equivalent to units of the i-theory (because of weight-sharing, all c_i are the same) but only when G is the translation group (in i-theory G is the full affine group for the first layers and can be a nongroup such as the transformation induced by rotations in depth).

Invariance and Deep Learning Networks

i-theory provides a general theory for DLNs while offering two extensions: (1) it ensures, within the same algorithm, invariances to other groups beyond translation and in an approximate way to certain nongroup transformations, and (2) it provides a way to learn arbitrary quasi invariances from implicitely supervised learning.

Invariance in 2-D and 3-D Vision

We assume that images as well as templates are in 2-D. This is the case if possible sources of 3-D information such as stereopsis and motion are eliminated. Interestingly, it seems that stereopsis does not facilitate recognition; this suggests that 3-D information, even when available, is not used by the human visual system (see [111]).

Relations to the Scattering Transform

There are connections between the scattering transform and i-theory but also several differences. There is no obvious correspondence between operations in the scattering transform and simple and complex cells in the ventral stream in contrast to convolutional networks and i-theory networks. i-theory provides an algorithm in which invariances are learned from implicitly supervised experience of transformations of a random set of objects/images; in the scattering transform invariances are hardwired. i-theory proves that Gabor-like templates are optimal for simultaneous invariance to scale and shift and that such invariance requires a multiresolution inverted truncated pyramid, which turns out to be reflected in the architecture of the visual cortex, starting with the eccentricity-dependent organization of the retina; in the scattering transform, Gabor wavelets are assumed at the start.

Explicit or Implicit Gating of Object Classes

The second stage of the recognition architecture consists of a large set of object class-specific modules of which probably the most important is the face network. It is natural to think that signals from lower areas should be gated, in order to route access only to the appropriate module. In fact, Tsao and Living Stone [112] postulated a gate mechanism for the network of face patches. The structure of the modules, however, suggests that the modules themselves provide a gating function even if their primary computational function is invariance. This is especially

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

clear in the case of the module associated with a single object, (the object class consists of a single object, as in the case of a paperclip). The input to the module is subject to dot products with each of the stored views of the object. If none matches well enough, the output of the module will be close to zero, effectively gating off the signal and switching off subsequent stages of processing.

Invariance to X and Estimation of X

Our description of i-theory focuses on the problem of recognition as estimating identity or category invariantly to a transformation (X) such as translation or scale or pose. Often however, the complementary problem, of estimating X, for instance, pose, is also important. Empirically a neural population of cells in IT is capable of supporting both representations; this multiplexed representation was shown in IT recordings [113] and model simulations [38]. As human observers, we are certainly able to estimate position, rotation, illumination of an object without eye movements. Different HW complex cells pooling over the same simple cells in different way—pooling over identities for each pose or pooling over pose for each identity—can provide the different types of information. Anselmi et al. [9, fig. 45] show simulations of recognizing a specific body invariantly to pose and estimating pose of a body invariantly to the identity.

PCA versus ICA

Independent component analysis (ICA) [114] and similar unsupervised mechanisms describe plasticity rules similar to the basic Oja flow. They can generate Gabor-like receptive fields, and they may not need the assumption of different sizes of Gaussian distributions of lateral geniculate nucleus (LGN) synapses. We used principal component analysis (PCA) simply because its properties are easier to analyze and should be indicative of the properties of similar Hebbian-like mechanisms.

Parsing a Scene

Full parsing of a scene cannot be done in a single feedforward processing step in the ventral stream. It requires task-dependent top-down control; in general, multiple fixations; and therefore observation times longer than ≈ 100 msec. This follows also from the limited high-resolution region of the inverted pyramid model of the visual system, which i-theory predicts as a consequence of simultaneous invariance to shift and scale. In any case, full parsing of a scene is beyond what a purely feedforward model can provide.

Feedforward and Feedback

We have reviewed a forward theory of recognition and some of the related evidence. i-theory does not address top-down or recurrent or horizontal

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

connectivity and their computational role. It makes it easier, however, to consider plausible hypotheses. The inverted pyramid architecture that follows from scale and position invariance requires for everyday vision a tight loop between different fixations in which an efficient control module drives eye movements by combining task requirements with memory access. Within a single fixation, however, the space-scale inverted pyramid cannot be shifted in space. A key function that can be controlled from the top is access to memory of lower-level signatures. For instance, a reverse hierarchy strategy may use the top-level signature to provide a high-level categorical scene interpretation and then to select appropriate routing of specific local low-level signatures to the memory and classification stage. In addition, the parameters of pooling could be controlled in a feedback mode, including the choice of which scales to use, depending on the results of classification or memory access. The most obvious limitation of feedforward architectures is recognition in clutter, and the most obvious way around the problem is the attentional masking of large parts of the image under top-down control. Further, a realistic implementation of the present theory requires top-down control signals and circuits, supervising learning, and possibly fetching signatures from different areas and at different locations in a task-dependent way. A simple idea is that signatures from all layers access the associative memory or classifier module and then control iterations in visual recognition and processing. Of course, at lower layers there are many signatures, each one in different complex cell layer locations, while at the top layer there are only a small number of signatures—in the limit only one. An even more interesting hypothesis is that backprojections update local signatures at lower levels, depending on the scene class currently detected at the top (an operation similar to the top-down pass of Borenstein and Ullman [115]). Thus the output of the feedforward pass may be used to retrieve labels and routines associated with the image. Backprojections may implement an attentional focus of processing to reduce clutter effects and also to run visual routines [38] at various levels of the hierarchy.

Motion Helps Learning Isolated Templates

Ideally templates and their transformations should be learned without clutter. It can be argued that if the background changes between transformed images of the same template, the averaging effect intrinsic to pooling will mostly average out the effect of clutter during the unsupervised learning stage. Although this is correct, and we have computer simulations that provide empirical support to the argument, it is interesting to speculate that motion could provide a simple way to eliminate most of the clutter. Sensitivity to motion is one of the earliest visual computations to appear in the course of evolution and one of the

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

most primitive. Stationary images on the retina tend to fade away. Detection of relative movement is a strong perceptual cue in primate vision as well as in insect vision, probably with similar normalization-like mechanisms [116, 117]. Motion induced by the transformation of a template may then serve two important roles:

- to bind together images of the same template while transforming: continuity of motion is implicitly used to ensure that identity is preserved
- to eliminate background and clutter by effectively using relative motion between the target object and the background.

The required mechanisms are probably available in the retina and early visual cortex.

Different Levels of Understanding

i-theory is at several different levels addressing the computational goal of the ventral stream; the algorithms used, down to the architecture of visual cortex; its hierarchical architecture; and the neural circuits underlying tuning of cells. This is unlike most other models or theories.

Finally, we should mention that some of the existing models between neuroscience and machine learning, such as HMAX [47, 50, 118] and other convolutional neural networks [2, 34, 119], are special cases of i-theory. Despite significant advances in sensory neuroscience over the last five decades, a true understanding of the basic functions of the ventral stream in visual cortex has proved elusive. Thus it is interesting that our theory follows from a novel hypothesis about the main computational function of the ventral stream: the representation of new objects/images in terms of a signature that is invariant to transformations learned during visual experience, thereby allowing recognition from very few labeled examples—in the limit, just one. This view of the cortex may also represent a novel theoretical framework for the next major challenge in learning theory beyond the supervised learning setting, which is now relatively mature: the problem of representation learning, formulated here as the unsupervised learning of invariant representations that significantly reduce the sample complexity of the supervised learning stage.

5.5 Ideas for Research

Following are some potential weaknesses of i-theory and some comments.

"The theory is too good to be true." One of the main problems of the theory is that it seems too elegant (in the sense of physics) for biology.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Backprojections are not taken into account, and they are an obvious feature of the cortical anatomy, which any real theory should explain. Backprojections and top-down controls are, however, implied by the present theory. Since feedforward architectures for recognition do not perform well in the presence of clutter, attentional masking may be a possible solution. Attentional focus driven by cortical backprojections will reduce clutter effects and activate visual routines at various levels of the hierarchy [38]. A Simple extension of the theory is an associative memory (storing signatures) that controls lower levels of the hierarchy via backprojections.

Subcortical projections, such as projections to and from the pulvinar, *are not predicted by the theory*. The present theory is still in the "cortical chauvinism" camp. Hopefully somebody will rescue it.

Cortical areas are organized in a series of layers with specific types of cells and corresponding arborizations and connectivities. *The theory does not say anything at this point about this level of the circuitry.*

Some directions for future research follow.

Associative Architecture for Retrieval

In past work on HMAX we assumed that hierarchical architectures perform preprocessing of an image to compute a vector (signature), which is then provided as input to a classifier. This view is extended by assuming that *signature vectors*, not only at the top of the hierarchy but at every complex cell level, are input to an associative memory. In this way a number of properties of the image (and associations) can be recalled. Parenthetically we note that old *associative memories* can be regarded as vector-valued classifiers.

- A neuron and its 10^3–10^4 synapses provide the basic computational operation: a *dot product*. The input I to the neuron gives the scalar $\langle I, t \rangle$ as the output, where t is the vector of synaptic weights.
- A set of K neurons (simple cells) computes the matrix operation $M^i f$, where

$$M^i = \begin{pmatrix} g_0 t^i \\ \cdots \\ g_K t^i \end{pmatrix}.$$

- Then the output of each complex cell c_i is the average over i of $|\langle M^i, I \rangle|$, which can be written as the dot product $c = e^T |\langle M^i, I \rangle|$, where $e = (1, \cdots, 1)^T$.
- The signature vector c is used to access an associative memory represented as a matrix A. Let us denote a specific vector c as c^j, and assume that the elements of the matrix M have stored by associating the j signature with a set

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

of properties given by the vector p^j: $M_{k,l} = \sum_j p^j_k c^j_l$. If c^n is noiselike, then $Mc^n \sim p^n$.

Retrieving from an Associative Memory: Optimal Sparse Encoding and Recall

There are estimates of optimal properties of codes for associative memories, including optimal sparseness (see [120, 121]). It would be interesting to connect these results to estimated capacity of visual memory [121].

Weak Labeling by Association of Video Frames

Assume that the top associative module associates images in a video that are contiguous in time (except when there are clear transitions). This idea (Kai Yu, personal communication) relies on smoothness in time to label via association. It is biological semisupervised learning, essentially identical to a Földiák-type rule. It is thus very much in tune with our proposal of the simple complex memory-based module for learning invariances to transformations and with ideas about an associative memory module at the very top.

Invariance and Perception

Other invariances in visual perception may be analyzed in a parallel way. An example is color constancy. Invariance to illumination (and color opponent cells) may emerge during development in a similar way as invariance to affine transformations. Thus we have a *color constancy* conjecture. The theory of chapter 1 should be able to learn invariance to illumination by observing during development transformations in the appearance of the same scene under changes of the illuminant—direction and spectral composition. A natural conjecture emerging from the approach in chapter 2 is that eigenvectors of the covariance matrix of such transformations of natural images may provide the spatial-chromatic tuning of different types of color opponent cells in V1 and other areas.

The idea that the key computational goal of the visual cortex is to learn and exploit invariances extends to other sensory modalities such as hearing of sounds and speech. It is tempting to think of music as an abstraction (in the sense of information compression and PCA) of the transformations of sounds. Classical (Western) music would then emerge from the transformations of human speech (the roots of Western classical music were based in the human voice—Gregorian chants).

Visual Concepts

- *Concept of Parallel Lines*. Consider an architecture using signatures. Assume it has learned sets of templates that guarantee invariance to all affine

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

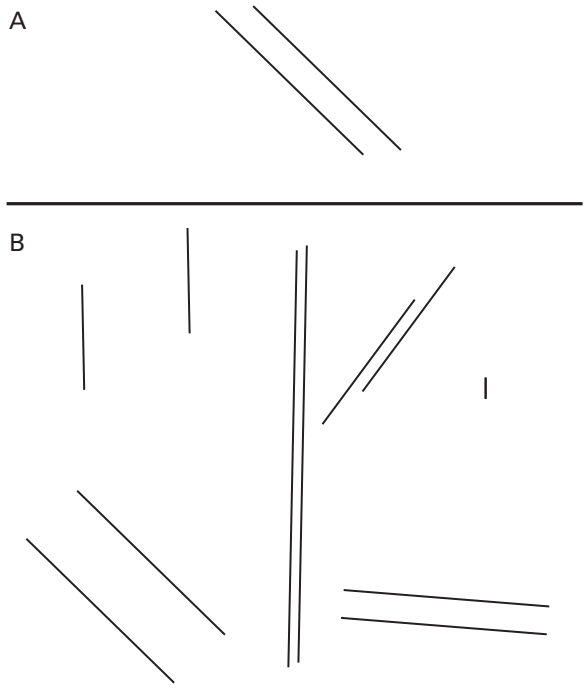

Figure 5.1

For a system that is invariant to affine transformations, a single training example A allows recognition of all other instances of parallel lines never seen before.

transformations. The claim is that *the architecture will appear to have learned the concept of parallel lines from a single specific example of two parallel lines* (see figure 5.1). According to the theorems presented here, the signature of the single image of the parallel lines will be invariant to affine transformations (within some range).

- *Number of items in an image*. A classifier that learns the number 5 in a way that is invariant to scale should be able to recognize five objects independent of class of objects.
- *Line drawings conjecture*. The memory-based module described here should be able to generalize from real images to line drawings when exposed to illumination-dependent transformations of images. This may need to happen at more than one level in the system, starting with the very first layer (e.g., V1). Generalizations with respect to recognition of objects invariant to shadows may also be possible.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Appendix

A.1 Invariant Representations and Bounds on Learning Rates: Sample Complexity

In this section we derive a mathematical proof of how an invariant representation can effectively reduce the sample complexity of the recognition task. In particular, we show that far fewer examples are needed to train a classifier for recognition (sample complexity) if we use an invariant representation. We start with the basic problem of learning how two sets of random variables x, y are probabilistically related. If their relation is described by the distribution $P(x, y)$, the problem is usually solved by minimizing the square loss

$$E(f) = \int (y - f(x))^2 P(x, y) dx\, dy.$$

However, the distribution $P(x, y)$ is unknown, and what can be minimized is only the empirical error of a set or random sampled points (x_i, y_i), $i = 1, \ldots, m$,

$$E_m(f) = \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2. \tag{A.1}$$

Let f_m be the minimizer in eq. (A.1). We would like to have that the probability that $E(f_m)$ and $E_m(f_m)$ differ significantly is very small,

$$Prob|E(f_m) - E_m(f_m) > \epsilon| < 1 - \delta, \quad \epsilon, \delta > 0. \tag{A.2}$$

This problem is related (see, e.g., [122]) to the calculation of the *covering number* of the function space (hypothesis space), $\mathcal{F}$, where the minimizer lives. More precisely, if the samples (x_i, y_i) are contained in a ball B of radius r, the covering number, $\mathcal{N}(\epsilon, B, \mathcal{F})$, is defined as the minimum number of $\epsilon-$ balls needed to cover B. It can be proved [122] that

$$\delta = 2\mathcal{N}(\mathcal{F}, \frac{\epsilon}{8}) e^{-\frac{m\epsilon^2}{8}}. \tag{A.3}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Inverting eq. (A.3), we have a bound on the number of samples, m, in terms of the covering number of the functional space of the possible minimizers.

More concretely, suppose the samples are taken from a sample space $\mathcal{X}$ of images, where the distribution $P(x, y)$ is defined. If, for example, we express the images in a basis, they can be viewed as vectors in $\mathbb{R}^d$ (e.g., discrete wavelet basis). The sample complexity of a learning rule depends on the covering number of the ball, $B \in \mathbb{R}^d$, that contains all the image distribution:

$$\mathcal{N}(\epsilon, B, \mathcal{X}) \sim \left(\frac{r}{\epsilon}\right)^d.$$

In the case of linear learning rules, the sample complexity is proportional to the *logarithm* of the covering number.

As a basic example, consider $\mathcal{X}$ to be d-dimensional and a hypothesis space of linear functions

$$f(I) = \langle w, I \rangle, \quad \forall I \in \mathcal{X}, w \in \mathcal{X},$$

so that the data representation is simply the identity. Then the ϵ-covering number of the set of linear functions with $\|w\| \leq 1$ is given by

$$N_\epsilon \sim \epsilon^{-d}.$$

If the input data lie in a subspace of dimension $s \leq d$, then the covering number of the space of linear functions becomes $N_\epsilon \sim \epsilon^{-s}$. In the next section we further comment on this example and provide an argument to illustrate the potential benefits of invariant representations.

A.1.1 Translation Group

Consider the simple example of a set of images of $p \times p$ pixels, and each containing an object within a (square) window of $k \times k$ pixels and surrounded by a uniform background. Imagine the object positions to be possibly anywhere in the image. Without loss of generality, fix the window centered in the middle of the image. Then it is easy to see that as soon as objects are translated so that they do not overlap, we get an orthogonal subspace. Then we see that there are $r^2 = (p/k)^2$ possible subspaces of dimension k^2, that is, the set of translated images can be seen as a distribution of vectors supported within a ball in $d = p^2$ dimensions. Following the discussion in the previous section, the best algorithm based on a linear hypothesis space will incur a sample complexity proportional to d. Assume now we have access to an oracle that can register each image so that each object occupies the centered position. In this case, the distribution of images is effectively supported within a ball in $s = k^2$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

dimensions, and the sample complexity is proportional to s rather than d. In other words, a linear learning algorithm would need

$$r^2 = d/s$$

fewer examples to achieve the same accuracy. The idea is that invariant representations can act as an invariance oracle and have the same impact on the sample complexity.

A.1.2 Scale and Translation: 1-D Affine Group

In the following we consider the case of a non-Abelian locally compact discrete group of transformations G and its associated wavelet transform (e.g., $g_{j,k} \in G,\ j,k \in \mathbb{Z}$ such that $g_{j,k}f(x) = 2^{j/2}f(2^j x - k)$, $f \in L^2(\mathbb{R})$). Let $I \in L^2(\mathbb{R})$ be an image, and $\bar{G}$ be a finite subset of G. We have the following.

Lemma A.1 Let $I \in \mathcal{X}^\Omega = L^2(\Omega)$ be an image belonging to the set of square integrable functions with support in $\Omega \subseteq \mathbb{R}$ and $W : L^2(\mathbb{R}) \to \mathbb{R}^{|G|}$, the discrete wavelet transform operator, where $|\cdot|$ indicates the cardinality. Let $\mathcal{N}^{\bar{G}I}(B, \epsilon)$ (B is a ball of radius r) be the covering number of the space of functions obtained by applying the group transformations $g \in \bar{G}$ to I. Let $m_{\bar{G}I}$ be its associated sample complexity. We have

$$m_{\text{inv}} = m_{\text{image}} \frac{|G_I|}{|\bar{G}G_I|} \tag{A.4}$$

with

$$|G_I| = |\text{supp}(W(I))|,$$

where "supp" indicates the support, and W is the discrete wavelet transform operator.

Proof: Note first that eq. (A.4) is well defined, since the function I has compact support in Ω, that is, $|G_I| < \infty$, $|\bar{G}| < \infty$ by hypothesis, and $|\bar{G}G_\Omega| < |\bar{G}||G_I|$.

Given the definitions in the theorem, in the case of a linear classifier, we have

$$\mathcal{N}^I(B, \epsilon) = \left(\frac{r}{\epsilon}\right)^{|G_I|}.$$

Consider all the $\bar{G}$-group shifted versions of I. Using the covariance property of the wavelet coefficients, we have that the number of nonzero coefficients is $|\bar{G}G_I|$, and therefore the associated covering number is

$$\mathcal{N}^{\bar{G}I}(B, \epsilon) = \left(\frac{r}{\epsilon}\right)^{|\bar{G}G_I|}.$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Thus, with the sample complexity $m \propto \ln(\mathcal{N})$, we have

$$m_{\text{image}} \propto |\bar{G}G_I|.$$

The sample complexity gain is in this case

$$\frac{m_{\text{inv}}}{m_{\text{image}}} = \frac{|G_I|}{|\bar{G}G_I|}.$$

A.2 One-Layer Architecture: Invariance and Selectivity

A.2.1 Equivalence between Orbits and Probability Distributions

The following is taken from Anselmi et al. [1].

Let $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{P}(\mathcal{X})$ be the space of probability measures on $\mathcal{X}$. Recall that for any compact group, the Haar measure is finite, so if appropriately normalized, it corresponds to a probability measure.

In the following we assume G to be Abelian and compact and the corresponding Haar measure to be normalized. The first step in our reasoning is the following definition.

Definition A.2 Representation via Orbit Probability For all $I \in \mathcal{X}$, define the random variable

$$Z_I : (G, dg) \to \mathcal{X}, \quad Z_I(g) = gI, \quad \forall g \in G,$$

with law

$$\rho_I(A) = (Z_x)\rho(A) = \int_{Z_I^{-1}(A)} dg,$$

for all measurable sets $A \subset \mathcal{X}$. Let

$$P : \mathcal{X} \to \mathcal{P}(\mathcal{X}), \quad P(I) = \rho_I, \quad \forall I \in \mathcal{X}.$$

The map P associates to each point a corresponding probability distribution. From definition A.2 we see that we are essentially viewing an orbit as a distribution of points, and mapping each point in one such distribution. Then we have the following result.

Theorem A.3 For all $I, I' \in \mathcal{X}$,

$$I \sim I' \quad \Leftrightarrow \quad P(I) = P(I'). \tag{A.5}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Proof: We first prove that $I \sim I' \Rightarrow \rho_I = \rho_{I'}$. Recalling that if $\mathcal{C}_c(\mathcal{I})$ is the set of continuous functions on $\mathcal{I}$ with compact support, ρ_I can be alternatively defined as the unique probability distribution such that

$$\int f(z)d\rho_I(z) = \int f(Z_I(g))dg, \quad \forall f \in \mathcal{C}_c(\mathcal{I}). \tag{A.6}$$

Therefore $\rho_I = \rho_{I'}$ if and only if for any $f \in \mathcal{C}_c(\mathcal{I})$, we have $\int_{\mathcal{G}} f(Z_I(g))dg = \int_{\mathcal{G}} f(Z_{I'}(g))dg$, which follows immediately from a change of variable and invariance of the Haar measure:

$$\int_{\mathcal{G}} f(Z_I(g))dg = \int_{\mathcal{G}} f(gI)dg = \int_{\mathcal{G}} f(gI')dg = \int_{\mathcal{G}} f(g\tilde{g}I)dg = \int_{\mathcal{G}} f(\hat{g}I)d\hat{g}.$$

To prove that $\rho_I = \rho_{I'} \Rightarrow I \sim I'$, note that $\rho_I(A) - \rho_{I'}(A) = 0$ for all measurable sets $A \subseteq \mathcal{X}$ implies in particular that the support of the probability distributions of I has non-null intersection on a set of non-zero measure. Since the support of the distributions $\rho_I, \rho_{I'}$ is exactly the orbits associated to I, I', respectively, then the orbits coincide, that is, $I \sim I'$.

This result shows that an invariant representation can be defined considering the probability distribution naturally associated to each orbit. However, its computational realization would require dealing with high-dimensional distributions. Indeed, we next show that the representation can be further developed to consider only probability distributions on the real line.

A.2.2 Cramér-Wold Theorem and Random Projections for Probability Distributions

Simple operations for neurons are (high-dimensional) inner products, $\langle \cdot, \cdot \rangle$, between inputs and stored templates, which are neural images. In this section we show how it is possible to give a good estimate of the (high-dimensional) probability distribution described in the previous section through a collection of one-dimensional probability distributions.

Classical results such as the Cramér-Wold theorem [123], ensure that a probability distribution P_I can be almost uniquely characterized by K one-dimensional probability distributions $P_{\langle I,t^k \rangle}$ induced by the (one-dimensional) results of projections $\langle I, t^k \rangle$, where t^k, $k = 1, \ldots, K$ are a set of randomly chosen images called templates. A probability function in d variables (the image dimensionality) induces a unique set of 1-D projections, which is discriminative; empirically, a small number of projections is usually sufficient to discriminate among a finite number of different probability distributions. Theorem A.5 says (informally) that an approximately invariant and unique signature of an image I can be obtained from the estimates of K 1-D probability distributions

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

$P_{\langle I,t^k\rangle}$ for $k = 1,\ldots,K$. The number K of projections needed to discriminate n orbits, induced by n images, up to precision ϵ (and with confidence $1-\delta^2$) is $K \geq \frac{2}{c\epsilon^2}\log\frac{n}{\delta}$, where c is a universal constant. Thus the discriminability question can be answered positively (up to ϵ) in terms of empirical estimates of the one-dimensional distributions $P_{\langle I,t^k\rangle}$ of projections of the image onto a finite number of templates $t^k,\ k = 1,\ldots,K$, under the action of the group.

A.2.3 Finite Random Projections Almost Discriminate among Different Probability Distributions

Given the discussion in Chapter 1, a *signature* may be associated to I by constructing a histogram approximation of P_I, but this would require dealing with high-dimensional histograms. The following classic theorem gives a way around this problem.

For a *template* $t \in \mathbb{S}(\mathbb{R}^d)$, where $\mathbb{S}(\mathbb{R}^d)$ is unit sphere in $\mathbb{R}^d$, let $I \mapsto \langle I,t\rangle$ be the associated projection. Moreover, let $P_{\langle I,t\rangle}$ be the distribution associated to the random variable $g \mapsto \langle gI,t\rangle$ (or equivalently, $g \mapsto \langle I, g^{-1}t\rangle$, if g is unitary). Let $\mathcal{E} = [t \in \mathbb{S}(\mathbb{R}^d),\ \text{s.t. } P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}]$.

Theorem A.4 (Cramér-Wold [123]) For any pair P, Q of probability distributions on $\mathbb{R}^d$, we have that $P = Q$ if and only if $\mathcal{E} = \mathbb{S}(\mathbb{R}^d)$.

In words, two probability distributions are equal if and only if their projections on any of the unit sphere directions are equal. This result can be equivalently stated as saying that the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 1 if and only if $P = Q$, and the probability of choosing t such that $P_{\langle I,t\rangle} = Q_{\langle I,t\rangle}$ is equal to 0 if and only if $P \neq Q$ (see [123, theorem 3.4]). The theorem suggests a way to define a metric on distributions (orbits) in terms of

$$d(P_I, P_{I'}) = \int d_0(P_{\langle I,t\rangle}, P_{\langle I',t\rangle})d\lambda(t), \quad \forall I, I' \in \mathcal{X},$$

where d_0 is any metric on one-dimensional probability distributions, and $d\lambda(t)$ is a distribution measure on the projections. Indeed, it is easy to check that d is a metric. In particular, note that in view of the Cramér Wold theorem, $d(P,Q) = 0$ if and only if $P = Q$. As mentioned, each one-dimensional distribution $P_{\langle I,t\rangle}$ can be approximated by a suitable histogram $\mu^t(I) = (\mu^t_h(I))_{h=1,\ldots,H} \in R^H$, so that in the limit in which the histogram approximation is accurate,

$$d(P_I, P_{I'}) \approx \int d_\mu(\mu^t(I), \mu^t(I'))d\lambda(t), \quad \forall I, I' \in \mathcal{X}, \tag{A.7}$$

where d_μ is a metric on histograms induced by d_0.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A natural question is whether there are situations in which a finite number of projections suffice to discriminate any two probability distributions, that is, $P_I \neq P'_I \Leftrightarrow d(P_I, P_{I'}) \neq 0$. Empirical results show that this is often the case with a small number of templates (see [125] and HMAX experiments). The problem of mathematically characterizing the situations in which a finite number of (one-dimensional) projections are sufficient is challenging. Here we provide a partial answer to this question.

We start by observing that the metric (A.7) can be approximated by uniformly sampling K templates and considering

$$\hat{d}_K(P_I, P_{I'}) = \frac{1}{K} \sum_{k=1}^{K} d_\mu(\mu^k(I), \mu^k(I')), \tag{A.8}$$

where $\mu^k = \mu^{t^k}$. The following result shows that a finite number K of templates is sufficient to obtain an approximation within a given precision ϵ. Toward this end, let

$$d_\mu(\mu^k(I), \mu^k(I')) = \left\| \mu^k(I) - \mu^k(I') \right\|_{\mathbb{R}^H}, \tag{A.9}$$

where $\|\cdot\|_{\mathbb{R}^H}$ is the Euclidean norm in $\mathbb{R}^H$. The following theorem holds.

Theorem A.5 Consider n images $\mathcal{X}_n$ in $\mathcal{X}$. Let $K \geq \frac{2}{c\epsilon^2} \log \frac{n}{\delta}$, where c is a universal constant. Then

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| \leq \epsilon, \tag{A.10}$$

with probability $1 - \delta^2$, for all $I, I' \in \mathcal{X}_n$.

Proof: The proof follows from an application of the Hoeffding inequality (see box A.1) and a union bound.

Fix $I, I' \in \mathcal{X}_n$. Define the real random variable $Z : \mathbb{S}(\mathbb{R}^d) \to \mathbb{R}$,

$$Z(t^k) = \left\| \mu^k(I) - \mu^k(I') \right\|_{\mathbb{R}^H}, \ \ k = 1, \ldots, K.$$

From the definitions it follows that $\|Z\| \leq c$, and $\mathbb{E}(Z) = d(P_I, P_{I'})$. Then the Hoeffding inequality implies

$$|d(P_I, P_{I'}) - \hat{d}_K(P_I, P_{I'})| = |\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}(Z) - Z(t^k)| \geq \epsilon,$$

with probability at most $e^{-c\epsilon^2 k}$. A union bound implies a result holding uniformly on $\mathcal{X}_n$; the probability becomes at most $n^2 e^{-c\epsilon^2 K}$. The desired result is obtained, noting that this probability is less than δ^2 as soon as $n^2 e^{-c\epsilon^2 K} < \delta^2$, that is, $K \geq \frac{2}{c\epsilon^2} \log \frac{n}{\delta}$.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Box A.1
Hoeffding Inequality

Suppose we have a set of variables X_i. It is intuitive that when we average some of them, we should normally get a result close to the expected value. Hoeffding quantifies normally and close. More formally, let $X_1, \cdots, X_N$ be a set of independent, independently distributed random variables such that $0 \leq X_i \leq 1$. We have

$$Pr\Big(|\frac{1}{N}\sum_{i=1}^{N} X_i - E(X)| > \epsilon\Big) \leq \delta = 2e^{-2N\epsilon^2}.$$

This result shows that the discriminability question can be answered in terms of empirical estimates of the one-dimensional distributions of projections of the image and transformations induced by the group on a number of templates t^k, $k = 1, \ldots, K$.

Theorem A.5 can be compared to a version of the Cramér-Wold theorem for discrete probability distributions. Heppes [125, theorem 1] shows that for a probability distribution consisting of k atoms in $\mathbb{R}^d$, at most $k+1$ directions ($d_1 = d_2 = \ldots = d_{k+1} = 1$) are enough to characterize the distribution and thus a finite (albeit large) number of one-dimensional projections.

The signature $\Sigma(I) = (\mu_1^1(I), \ldots, \mu_H^K(I))$ is obviously invariant (and unique), since it is associated to an image and all its transformations (an orbit). Each component of the signature is also invariant; it corresponds to a group average. Indeed, each measurement can be defined as

$$\mu_h^k(I) = \frac{1}{|G|}\sum_{g \in G} \eta_h\Big(\Big\langle gI, t^k \Big\rangle\Big), \tag{A.11}$$

for G finite group, or equivalently,

$$\mu_h^k(I) = \int_G dg\ \eta_h\Big(\Big\langle gI, t^k \Big\rangle\Big) = \int_G dg\ \eta_h\Big(\Big\langle I, g^{-1}t^k \Big\rangle\Big), \tag{A.12}$$

when G is a (locally) compact group. Here, the nonlinearity η_h is chosen to define a histogram approximation. Then it is clear that from the properties of the Haar measure we have

$$\mu_h^k(\bar{g}I) = \mu_h^k(I), \quad \forall \bar{g} \in G, I \in \mathcal{X}. \tag{A.13}$$

This is a key property of the signature and of our entire model.

A.2.4 Number of Templates Depends on Pooling Size

Let us consider how a finite number of projections suffice to discriminate the distributions generated by the subset $|G|$ of a finite, discrete group. We assume

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

that the set of images (of dimensionality d, where d is the number of pixels) is also finite of cardinality n. Though not necessary for the main results of this section, we may also assume that the images themselves can assume a finite set of discrete (vector) values (b bits per pixel) so that the associated distribution $P_G(I)$ consists of $b^d \times |G|$ atoms.

Instead of "bits per pixel" a different representation can be used, such as bits per wavelet coefficient.

We think here of the group acting on the images, while the templates $t^k, k = 1, \cdots, K$ are fixed and not transformed. Later, once we have results connecting distributions on images induced by the group with one-dimensional distributions of projections of the images on the templates, we can think of the transformations acting on the templates instead of the images.

Extension of Johnson-Lindenstrauss (J-L)lemma to discrete distributions. Recall the Johnson-Lindenstrauss lemma.

Proposition A.6 For any set V of n points in $\mathbb{R}^d$, there exists a map $P : \mathbb{R}^d \to \mathbb{R}^K$ such that for all $u, v \in V$,

$$(1 - \epsilon) \parallel u - v \parallel \leq \parallel Pu - Pv \parallel \leq (1 + \epsilon) \parallel u - v \parallel,$$

where the map P is a random projection on $\mathbb{R}^K$ and

$$K \geq C(\epsilon)\ln(n), \quad C(\epsilon) = 4C_0\epsilon^{-2}.$$

Suppose u and v are two images in $\mathbb{R}^d$, and consider $P_G(u)$ and $P_G(v)$, generated by the action of G on u and v, respectively. Each distribution has G atoms (with possible but unlikely multiplicity) supported in $\mathbb{R}^d$. The key observation is that $P_G(u)$ and $P_G(v)$ are equivalent to the orbits of u and v, respectively, and thus contain the same sample points (e.g., images) irrespective of the order. If there is a point in $P_G(u)$ that is ϵ-different from any other point in $P_G(v)$, the two distributions are different. The J-L projections provide the required accuracy if $K \geq C_0\epsilon^{-2}\ln(n|G|)$. This is because in the J-L proof the Hoeffding inequality (implies the result in proposition (A.6)) for a specific u and v with probability at least $1 - e^{-c\epsilon^2 k}$. If we require this to hold uniformly for all $|G|$ elements of an orbit and for all $n(n-1)$ possible pairs of orbits, the number of required templates must satisfy $K \geq C_0\epsilon^{-2}\ln(n(n-1)|G|)$. The last step is to show that this implies relations between the multidimensional distributions $P_G(u), P_G(v)$ and the one-dimensional distributions $P_G(\langle u, t_k\rangle), P_G(\langle v, t_k\rangle)$. The correspondence rests on the discreteness of the distributions and the fact that the J-L lemma result implies that if u and v are very close (that is, $||u - v|| \leq \eta$), their projections $P(u)$ and $P(v)$ are very close in every norm, in particular componentwise (that is, $\max_k |P(u)_k - P(v)_k| \leq \eta$).

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

In particular, if each of the $|G|$ $g_i u$ and $g_j v$ are close to each other, it follows that the K distributions of the projections are also close, for instance, in the "sup" norm (and vice versa). If any of the distributions is not close, it means that at least one of the $g_i u$ is not close to any of the $g_j v$, and this in turn implies that the $P_G(u)$ and $P_G(v)$ are not close. The distributions can be computed from the components of Pu in terms of their moments or directly, since the number of atoms in each one-dimensional distribution is finite and equal to $|G|$.

Since if two orbits are different none of the $|G|$ elements can be the same, the conditions on the bounds may be relaxed. Somewhat counterintuitively, the comparison of two orbits should have higher confidence than the comparison of two points.

To give some perspective, consider an old-fashioned setup with the discrete distribution $P_G(u)$ and its moment-generating function $M_u(t) = \sum_{j=1}^{|G|} e^{\langle t, g_j u \rangle} p(g_j u)$. Choose $|G|$ appropriate templates t_i. Call m the $|G|$ dimensional vector of the moments and a the vector of probabilities. Then

$$m = Ma,$$

where M is the $|G| \times |G|$ matrix of $e^{\langle t_i, x_j \rangle}$. The equation can be solved for the probabilities a provided M^{-1} exists. This suggests that from the empirical moment-generating function it is possible to recover exactly the group orbits with a number of templates K in the order of $|G|$. We can summarize this reasoning in the following:

Theorem A.7 Let $\mathcal{X}_n$, $|\mathcal{X}_n| = n$, a finite subset of vectors from the vector space $\mathcal{X} = \mathbb{R}^d$, and let G be a finite discrete group. The probability distributions associated to any pair of orbits are ϵ-different iff any pair of the orbit points is ϵ-different. A similar result holds if we consider instead of the orbits their projections onto $K \propto \ln(n(n-1)|G|)$ random vectors and the associated probability distributions.

Proof: $\Leftarrow$ Let us start by considering a pair of vectors of $\mathcal{X}_n$, I, I' and their orbits (the generalization to n orbits is easy). Suppose the two orbits coincide up to an ϵ factor. This means that for the ordered $|G|$-couples of vectors ($v_i = g_i I \in O_I$, $u_i = g_i I' \in O_{I'}$) we have

$$\sup_j |(v_i)_j - (u_i)_j| \leq \epsilon, \ \forall i = 1, \ldots, |G|, \ j = 1, \ldots, d. \tag{A.14}$$

The condition in (A.14) guarantees that

$$KS(P_I, P_{I'}) \leq \frac{d|G|\epsilon}{d|G|} = \epsilon,$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

where KS is the Kolmogorov-Smirnov distance between the cumulative distribution functions of the orbits of I and I', calculated as

$$KS(P_I, P_{I'}) = \sup_i |\mathrm{cdf}_i(P_I) - \mathrm{cdf}_i(P_{I'})|,$$

where with cdf_i indicates the i^{th} component of the cdf histogram plotting the distribution of the vector values. In fact, the maximum difference between the cdfs will occur in the case when all vectors have the same entries; in that case, we will have an accumulation error from all the differences of $d\epsilon|G|$, which simplifies with the cdf normalization factor $1/d|G|$.

$\Rightarrow$ Conversely, if the probability distributions associated to two orbits differ by an ϵ factor in the orbit points, any couple of ordered points in the orbits can maximally differ by the same factor.

This reasoning can be repeated if we consider projections of the orbits onto random vectors and the associated probability distributions. By the J-L lemma we can distinguish $2|G|$ vectors (number of elements of the two orbits) with $K \propto \ln(2|G|)$ projections. If we want this to hold uniformly over $\mathcal{X}_n$, we have to consider all possible orbit couples, and therefore the number of required projections will be $K \propto \ln(n(n-1)|G|)$. We want to prove that this is enough for the $n(n-1)$ associated probability distributions to be distinguished. Consider the projections $\langle g_i I, t\rangle$, $I \in \mathcal{X}_n$, and construct the associated probability distributions. The same reasoning can be applied: instead of the orbit elements u_i, v_i, we have their projections $\langle u_i, t\rangle, \langle v_i, t\rangle$ and the associated probability distributions are $P_{\langle I,t\rangle}, P_{\langle I',t\rangle}$.

A.2.5 Partially Observable Groups

A.2.5.1 Invariance for Partially Observable Groups Implies Localization

In this section we analyze how the invariance property described in eq. (A.13) of the signature defined in eqs. (A.11) and (A.12) changes when we do not have access to the whole set of the transformations (orbits) but to part of it. In particular, we show how invariance is linked to localization of the dot product $\langle I, gt\rangle$ in the group transformation.

Since the group is only partially observable, we introduce the notion of *partial invariance* for images and transformations G_0 that are within the observation window. Partial invariance is defined in terms of invariance of

$$\mu_h^k(I) = \frac{1}{V_0} \int_{G_0} dg\ \eta_h\Big(\Big\langle gI, t^k \Big\rangle\Big). \tag{A.15}$$

We recall that when gI and t^k do not share any common support on the plane, or I and t are uncorrelated, then $\langle gI, t^k\rangle = 0$. The following theorem, where

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

G_0 corresponds to the pooling range states, is a sufficient condition for partial invariance in the case of a locally compact group (see also figure A.1).

Theorem A.8 (Localization and Invariance) Let $I, t \in H$, be a Hilbert space, $\eta_h : \mathbb{R} \to \mathbb{R}^+$ a set of bijective (positive) functions, and G a locally compact group. Let $G_0 \subseteq G$, and suppose supp $(\langle gI, t^k \rangle) \subseteq G_0$. Then for any given $\bar{g} \in G$, $t^k, I \in \mathcal{X}$, the following conditions hold:

$$\begin{aligned} &\left\langle gI, t^k \right\rangle = 0, \ \forall g \in G/(G_0 \cap \bar{g}G_0) \\ &\text{or equivalently,} \qquad\qquad \Rightarrow \ \mu_h^k(I) = \mu_h^k(\bar{g}I) \\ &\left\langle gI, t^k \right\rangle \neq 0, \ \forall g \in G_0 \cap \bar{g}G_0. \end{aligned} \tag{A.16}$$

Proof: To prove the implication, note that if $\left\langle gI, t^k \right\rangle = 0, \ \forall g \in G/(G_0 \cap \bar{g}G_0)$, with $G_0 \Delta \bar{g}G_0 \subseteq G/(G_0 \cap \bar{g}G_0)$ (Δ is the symbol for symmetric difference $A \Delta B = (A \cup B)/(A \cap B)$ A, B sets), we have

$$\begin{aligned} 0 &= \int_{G/(G_0 \cap \bar{g}G_0)} dg \ \eta_h\Big(\left\langle gI, t^k \right\rangle\Big) \\ &= \int_{G_0 \Delta \bar{g}G_0} dg \ \eta_h\Big(\left\langle gI, t^k \right\rangle\Big) \\ &\geq \Big| \int_{G_0} dg \ \Big(\eta_h\Big(\left\langle gI, t^k \right\rangle\Big) - \eta_h\Big(\left\langle g\bar{g}I, t^k \right\rangle\Big)\Big)\Big|. \end{aligned} \tag{A.17}$$

The second equality is true, since, being η_h positive, the fact that the integral is zero implies $\left\langle gI, t^k \right\rangle = 0 \ \forall g \in G/(G_0 \cap \bar{g}G_0)$ (and therefore in particular $\forall g \in G_0 \Delta \bar{g}G_0$). The right hand side of the inequality being positive, we have

$$\Big| \int_{G_0} dg \ \Big(\eta_h\Big(\left\langle gI, t^k \right\rangle\Big) - \eta_h\Big(\left\langle g\bar{g}I, t^k \right\rangle\Big)\Big)\Big| = 0, \tag{A.18}$$

that is, $\mu_h^k(I) = \mu_h^k(\bar{g}I)$ (see figure A.1 for a graphical representation).

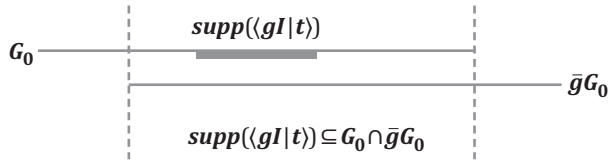

Figure A.1

A sufficient condition for invariance for locally compact groups. If the support of $\langle gI, t \rangle$ is sufficiently localized, it will be completely contained in the pooling interval even if the image is group shifted or, equivalently (as shown), if the pooling interval is group shifted by the same amount.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Eq. (A.16) describes a *localization* condition on the inner product of the transformed image and the template. The preceding result naturally raises the question of whether the localization condition is also necessary for invariance. Clearly, this would be the case if eq. (A.17) could be turned into an equality, that is,

$$\begin{aligned} &\int_{G_0 \Delta \bar{g} G_0} dg\ \eta_h\big(\big\langle gI, t^k \big\rangle\big) \\ &= |\int_{G_0} dg\ \Big(\eta_h\big(\big\langle gI, t^k \big\rangle\big) - \eta_h\big(\big\langle g\bar{g}I, t^k \big\rangle\big)\Big)| \\ &= |\mu_h^k(I) - \mu_h^k(\bar{g}I)|. \end{aligned} \tag{A.19}$$

Indeed, in this case, if $\mu_h^k(I) - \mu_h^k(\bar{g}I) = 0$, and we further assume the natural condition $\langle gI, t^k \rangle \neq 0$ if and only if $g \in G_0$, then the localization condition (1.8) would be necessary, since η_h is a positive bijective function.

Invariance, Localization, and Wavelets

The conditions equivalent to optimal translation and scale invariance, which correspond to maximum localization in space and frequency, cannot be simultaneously satisfied because of the classic *uncertainty principle*: if a function $t(x)$ is essentially zero outside an interval of length Δx and its Fourier transform $\hat{I}(\omega)$ is essentially zero outside an interval of length $\Delta\omega$, then

$$\Delta x \cdot \Delta\omega \geq 1. \tag{A.20}$$

In other words, a function and its Fourier transform cannot both be highly concentrated. Interestingly, for our setup the uncertainty principle also applies to sequences (see [25]).

It is well known that the equality sign in the uncertainty principle above is achieved by Gabor functions (see [24]) of the form

$$\psi_{x_0,\omega_0}(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x},\ \sigma_x \in \mathbb{R}^+,\ \omega_0 \in \mathbb{R}. \tag{A.21}$$

The uncertainty principle leads to the concept of optimal localization instead of exact localization. In a similar way, it is natural to relax our definition of strict invariance (e.g., $\mu_h^k(I) = \mu_h^k(g'I)$) and to introduce ϵ-invariance as $|\mu_h^k(I) - \mu_h^k(g'I)| \leq \epsilon$. If we suppose, for instance, the localization condition

$$\langle T_x I, t\rangle = e^{-\frac{x^2}{\sigma_x^2}}, \quad \langle D_s I, t\rangle = e^{-\frac{s^2}{\sigma_s^2}},\ \sigma_x, \sigma_s \in \mathbb{R}, \tag{A.22}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

we have

$$|\mu_h^k(T_{\bar{x}}I) - \mu_h^k(I)| = \frac{1}{2}\sqrt{\sigma_x}\Big(\text{erf}\big([-b,b]\Delta[-b+\bar{x},b+\bar{x}]\big)\Big),$$
$$|\mu_h^k(D_{\bar{s}}I) - \mu_h^k(I)| = \frac{1}{2}\sqrt{\sigma_s}\Big(\text{erf}\big([-1/S,S]\Delta[\bar{s}/S,S\bar{s}]\big)\Big),$$

where "erf" is the error function. The differences, with an opportune choice of the localization ranges σ_s, σ_x, can be made as small as desired.

We end with a conjecture: the optimal ϵ-invariance is satisfied by templates with noncompact support that decay exponentially, such as a Gaussian or a Gabor wavelet. We can then speak of *optimal invariance* meaning optimal ϵ-invariance. This reasoning leads to the following theorem:

Theorem A.9 Assume invariants are computed from pooling within a pooling window with a set of linear filters. Then the optimal templates (e.g., filters) for maximum simultaneous invariance to translation and scale are Gabor functions

$$t(x) = e^{-\frac{x^2}{2\sigma^2}} e^{i\omega_0 x}. \tag{A.23}$$

Summarizing, we showed why Gabor templates are implied if we want optimal invariance in the case of partially observable groups of dilations and translations.

- The Gabor function $\psi_{x_0,\omega_0}(x)$ corresponds to a *Heisenberg box*, which has an x-spread $\sigma_x^2 = \int x^2|g(x)|dx$ and an ω spread $\sigma_\omega^2 = \int \omega^2|\hat{g}(\omega)|d\omega$ with area $\sigma_x\sigma_\omega$. Gabor wavelets arise under the action on $\psi(x)$ of the translation and scaling groups as follows. The function $\psi(x)$, as defined, is zero-mean and normalized,

  $$\int \psi(x)dx = 0 \tag{A.24}$$

  and

  $$\|\psi(x)\| = 1. \tag{A.25}$$

  A family of Gabor wavelets is obtained by translating and scaling ψ:

  $$\psi_{u,s}(x) = \frac{1}{s^{\frac{1}{2}}}\psi(\frac{x-u}{s}). \tag{A.26}$$

  Under certain conditions (in particular, the Heisenberg boxes associated with each wavelet must together cover the space-frequency plane) the Gabor wavelet family becomes a Gabor wavelet frame.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

- Optimal self-localization of the templates (which follows from localization), when valid simultaneously for space and scale, is also equivalent to Gabor wavelets. If they are a frame, full information can be preserved in an optimal quasi-invariant way.

A.2.6 Approximate Invariance to Nongroup Transformations

In the previous section we analyzed the relation between localization and invariance in the case of group transformations. By relaxing the requirement of exact invariance and exact localization we show how the same strategy for computing invariants can still be applied even in the case of nongroup transformations if certain localization properties of $\langle TI, t\rangle$ hold, where T is a smooth transformation.

We first notice that the localization condition of theorems A.8 and A.9, when relaxed to approximate localization, takes the form (e.g., for $1-D$ translations group) $\left\langle I, T_x t^k \right\rangle < \delta \quad \forall x \text{ s.t. } |x| > a$, where δ is small in the order of $1/\sqrt{d}$ (n is the dimension of the space) and $\left\langle gI, t^k \right\rangle \approx 1 \quad \forall x \text{ s.t. } |x| < a$.

We call this property *sparsity of I in the dictionary t^k under G.* This condition can be satisfied by templates that are similar to images in the set and are sufficiently rich to be incoherent for small transformations. Note that from the preceding reasoning the sparsity of I in t^k under G is expected to improve with increasing d and with noiselike encoding of I and t^k by the architecture.

Another important property of sparsity of I in t^k (in addition to allowing local approximate invariance to arbitrary transformations) is *clutter tolerance* in the sense that if n_1, n_2 are additive uncorrelated spatial noisy clutter $\left\langle I + n_1, gt^k + n_2 \right\rangle \approx \langle I, gt\rangle$.

Interestingly, the *sparsity condition under the group* is related to associative memories, for instance, of the holographic type (see [121, 127]). If the sparsity condition holds only for $I = t^k$ and for very small set of $g \in G$ (if it has the form $\left\langle I, gt^k \right\rangle = \delta(g)\delta_{I,t^k}$) (this implies strict memory-based recognition (see noninterpolating look-up table in [52]) with inability to generalize beyond stored templates or views.

While the first regime — exact (or ϵ-) invariance for generic images, yielding universal Gabor templates — applies to the first layer of the hierarchy, this second regime (sparsity) — approximate invariance for a class of images, yielding class-specific templates — is important for dealing with nongroup transformations at the top levels of the hierarchy, where receptive fields may be as large as the visual field.

Several interesting transformations do not have the group structure, for instance, the change of expression of a face or the change of pose of a body. We

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

show here that approximate invariance to transformations that are not groups can be obtained if the approximate localization condition holds, and if the transformation can be locally approximated by a linear transformation, for instance, a combination of translations, rotations, and nonhomogeneous scalings, which corresponds to a locally compact group admitting a Haar measure.

As an example, consider the transformation induced on the image plane by rotation in depth of a face. It can be decomposed into piecewise linear approximations around a small number of key templates, each one corresponding to a specific 3-*D* rotation of a template face. Each key template corresponds to a complex cell containing as simple cells a number of observed transformations of the key template within a small range of rotations. Each key template corresponds to a different signature that is invariant only for rotations around its center. Notice that the form of the linear approximation or the number of key templates needed does not affect the algorithm or its implementation. The templates learned are used in the standard dot-product-and-pooling module. The choice of the key templates, each one corresponding to a complex cell and thus to a signature component, is not critical as long as there are enough of them. For one-parameter groups, the key templates correspond to the knots of a piecewise linear spline approximation. Optimal placement of the centers, if desired, is a separate problem that we leave aside for now.

A.3 Multilayer Architecture: Invariance, Covariance, and Selectivity

A.3.1 Recursive Definition of Simple and Complex Responses

So far we have studied the invariance, uniqueness, and stability properties of signatures, both in the case when a whole group of transformations is observable (see eq. (1.2) in chapter 1), and in the case in which it is only partially observable (see eqs. (A.11) and (A.12)). We now discuss how the preceding ideas can be iterated to define a multilayer architecture.

Consider first the case when G is finite. Given a subset $G_0 \subset G$, we can associate a window gG_0 to each $g \in G$. Then we can use eq. (1.5) to define for each window a signature $\Sigma(I)(g)$ given by the measurements

$$\mu_h^k(I)(g) = \frac{1}{|G_0|} \sum_{\bar{g} \in gG_0} \eta_h\Big(\big\langle I, \bar{g}t^k \big\rangle\Big).$$

Note that the averaging in the integral is done on transformed templates, not on transformed images. We keep this form as the definition of signature. For fixed h, k a set of measurements corresponding to different windows can be seen as a $|G|$-dimensional vector. A signature $\Sigma(I)$ for the whole image

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

is obtained as *a signature of signatures*, that is, a collection of signatures $(\Sigma(I)(g_1), \ldots, \Sigma(I)(g_{|G|})$ associated to each window.

Since we assume that the output of each module is made zero-mean and normalized before further processing at the next layer, *conservation of information from one layer to the next requires saving the mean and the norm* at the output of each module at each level of the hierarchy. We conjecture that the neural image at the first layer is uniquely represented by the final signature at the top of the hierarchy and the means and norms at each layer (see [1]).

This discussion can be extended to continuous (locally compact) groups considering,

$$\mu_h^k(I)(g) = \frac{1}{V_0} \int_{gG_0} d\bar{g} \eta_h\Big(\Big\langle I, \bar{g}t^k \Big\rangle\Big), \quad V_0 = \int_{G_0} d\bar{g},$$

where, for fixed h,k, $\mu_h^k(I) : G \to \mathbb{R}$ can now be seen as a function on the group. In particular, if we denote by $K_0 : G \to \mathbb{R}$ the indicator function on G_0, we can write

$$\mu_h^k(I)(g) = \frac{1}{V_0} \int_G d\bar{g} K_0(\bar{g}^{-1}g) \eta_h\Big(\Big\langle I, \bar{g}t^k \Big\rangle\Big).$$

The signature for an image can again be seen as a collection of signatures corresponding to different windows, but in this case it is a function $\Sigma(I) : G \to \mathbb{R}^{HK}$, where $\Sigma(I)(g) \in \mathbb{R}^{HK}$, is a signature corresponding to the window G_0 centered at $g \in G$.

This construction can be iterated to define a hierarchy of signatures. Consider a sequence $G_1 \subset G_2, \ldots, \subset G_L = G$. For $q : G \to \mathbb{R}^p$, $p \in \mathbb{N}$ with an abuse of notation, we let $gq(\bar{g}) = q(g^{-1}\bar{g})$. Then we can consider the following construction.

We call *complex operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $\mu_\ell(I) : G \to \mathbb{R}^{HK}$,

$$\mu_\ell^{h,k}(I)(g) = \frac{1}{|G_\ell|} \sum_{\bar{g} \in gG_\ell} \eta_h\Big(v_\ell^k(I)(\bar{g})\Big), \tag{A.27}$$

and *simple operator* at layer ℓ the operator that maps an image $I \in \mathcal{X}$ to a function $v_\ell(I) : G \to \mathbb{R}^K$,

$$\mu_\ell^k(I)(g) = \Big\langle \mu_{\ell-1}(I), gt_\ell^k \Big\rangle, \tag{A.28}$$

with t_ℓ^k the kth template at layer ℓ and $\mu_0(I) = I$. Several comments are in order:

- Besides the first layer, the inner product defining the simple cell operator is that in $L^2(G) = \{q : G \to \mathbb{R}^{HK}, \mid \int dg |q(g)|^2 < \infty\}$.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

- The index ℓ corresponds to different layers, corresponding to different subsets G_ℓ.
- At each layer a (finite) set of templates $\mathcal{T}_\ell = (t_\ell^1, \ldots, t_\ell^K) \subset L^2(G)$ ($\mathcal{T}_0 \subset \mathcal{X}$) is assumed to be available. For simplicity, we assume that $|\mathcal{T}_\ell| = K$ for all $\ell = 1, \ldots, L$. The templates at layer ℓ can be thought of as *compactly supported functions*, with support much smaller than the corresponding set G_ℓ. Typically templates can be seen as image patches in the space of complex operator responses, that is, $t_\ell = \mu_{\ell-1}(\bar{t})$ for some $\bar{t} \in \mathcal{X}$.
- Similarly we assume (for simplicity in dealing with indexes) that the number of nonlinearities η_h, considered at every layer, is the same.

The extension to continuous' (locally compact) groups is straightforward. We express it in the following definition.

Definition A.10 Simple and Complex Responses For $\ell = 1, \ldots, L$, let $\mathcal{T}_\ell = (t_\ell^1, \ldots, t_\ell^K) \subset L^2(G)$ (and $\mathcal{T}_0 \subset \mathcal{X}$) be a sequence of template sets. The complex operator at layer ℓ maps an image $I \in \mathcal{X}$ to a function $\mu_\ell(I) : G \to \mathbb{R}^{HK}$. In components

$$\mu_\ell^{h,k}(I)(g) = \frac{1}{V_\ell} \int d\bar{g} K_\ell(\bar{g}^{-1} g) \eta_h \left(\nu_\ell^k(I)(\bar{g}) \right), \; g \in G, \tag{A.29}$$

where K_ℓ is the indicator function on G_ℓ, $V_\ell = \int_{G_\ell} d\bar{g}$ and

$$\nu_\ell^k(I)(g) = \left\langle \mu_{\ell-1}(I), g t_\ell^k \right\rangle, \quad g \in G, \tag{A.30}$$

($\mu_0(I) = I$) is the simple operator at layer ℓ that maps an image $I \in \mathcal{X}$ to a function $\nu_\ell(I) : G \to \mathbb{R}^K$.

Remark A.11 Note that eq. (A.29) can be written as

$$\mu_\ell^{h,k}(I) = K_\ell * \eta_h(\nu_\ell^k(I)), \tag{A.31}$$

where $*$ is the group convolution.

A.3.2 Inheriting Transformations: Covariance

We call the map Σ covariant iff

$$\Sigma(gI) = g^{-1} \Sigma(I), \quad \forall g \in G, I \in \mathcal{X},$$

when g acting on Σ is properly defined. In the following we show the covariance property for the $\mu_1^{h,k}$ response (figure A.2). Inductive reasoning then can be applied for higher-order responses. *We assume that the architecture*

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Figure A.2
Covariance: the response for an image I at position g is equal to the response of the group-shifted image at the shifted position.

is isotropic in the relevant covariance dimension (this implies that all the modules in each layer should be identical with identical templates) and that there is a continuum of modules in each layer. The following proposition proves the covariance property of the signature defined in eq. (A.29).

Proposition A.12 Let G be a locally compact group and $\bar{g} \in G$. Let $\mu_1^{h,k}$ be as defined in (A.29). Then $\mu_1^{h,k}(\tilde{g}I)(g) = \mu_1^{h,k}(I)(\tilde{g}^{-1}g)$.

Proof: Using eq. (A.29), we have

$$\begin{aligned}
\mu_1^{h,k}(\tilde{g}I)(g) &= \frac{1}{V_1}\int_G d\bar{g} K_1(\bar{g}^{-1}g)\eta_h\left(\left\langle \tilde{g}I, \bar{g}t^k\right\rangle\right) \\
&= \frac{1}{V_1}\int_G d\bar{g} K_1(\bar{g}^{-1}g)\eta_h\left(\left\langle I, \tilde{g}^{-1}\bar{g}t^k\right\rangle\right) \\
&= \frac{1}{V_1}\int_G d\hat{g} K_1(\hat{g}^{-1}\tilde{g}^{-1}g)\eta_h\left(\left\langle I, \hat{g}t^k\right\rangle\right) \\
&= \mu_1^{h,k}(I)(\tilde{g}^{-1}g),
\end{aligned}$$

where in the third line we used the change of variable $\hat{g} = \tilde{g}^{-1}\bar{g}$ and the invariance of the Haar measure.

Remark A.13 The covariance property described in proposition A.12 can be stated equivalently as $\mu_1^{h,k}(I)(g) = \mu_1^{h,k}(\bar{g}I)(\bar{g}g)$. This expression has a intuitive meaning, as shown in figure A.2.

Remark A.14 With respect to the range of invariance, the following property holds for multilayer architectures in which the output of a layer is defined as covariant if it transforms in the same way as the input. For a given transformation of an image or part of it, the signature from complex cells at a certain layer is either invariant or covariant with respect to the group of transformations. If it is covariant, there will be a higher layer in the network at which it is invariant (see theorem A.16), assuming that the image is contained in the

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

visual field. This property gives a *stratification* of ranges of invariance in the ventral stream: invariances should appear in a sequential order, meaning that smaller transformations will be invariant before larger ones, in earlier layers of the hierarchy (see [128]).

Remark A.15 *Covariance and wavelets*. The result provend in Theorem A.9 shows how optimal invariance of complex cells for scale and translation transformations leads to Gabor wavelet templates, and thus simple cells operate a Gabor wavelet transform of the image. However, the necessity of the wavelet transform (not necessarily Gabor) can be also proved from the very general requirements of linearity, covariance, and energy conservation of the representation,[129, p 41]. A related approach was taken by Stevens in [130], who showed that (1) the wavelet transform is covariant with the similitude group, and (2) the wavelet transform follows from the requirement of covariance (rather than invariance) for the simple cells.

A.3.2.1 Invariance in Hierarchical Architectures

We now find the conditions under which the functions μ_ℓ are locally invariant, that is, invariant within the restricted range of the pooling. We further prove that the range of invariance increases from layer to layer in the hierarchical architecture. The fact that for an image, in general, no more global invariance is guaranteed allows a novel definition of parts of an image.

The local invariance conditions are a simple reformulation of theorem A.8 in the context of a hierarchical architecture. In the following, for simplicity, we suppose that at each layer we only have a template t and a nonlinear function η.

Proposition A.16 (Invariance and Localization: Hierarchy) Let $I, t \in H$,be a Hilbert space, $\eta : \mathbb{R} \to \mathbb{R}^+$ a bijective (positive) function, and G a locally compact group. Let $G_\ell \subseteq G$, and suppose $\text{supp}(\langle g\mu_{\ell-1}(I), t\rangle) \subseteq G_\ell$. Then for any given $\bar{g} \in G$,

$$\begin{aligned} \mu_\ell(I) = \mu_\ell(\bar{g}I) \;\Leftrightarrow\; & \langle g\mu_{\ell-1}(I), t\rangle = 0,\ g \in G/(G_\ell \cap \bar{g}G_\ell), \\ & \langle g\mu_{\ell-1}(I), t\rangle \neq 0,\ g \in G_\ell \cap \bar{g}G_\ell. \end{aligned} \tag{A.32}$$

Proof: The proof follows the reasoning in theorem A.8 with I substituted by $\mu_{\ell-1}(I)$ using the covariance property $\mu_{\ell-1}(gI) = g\mu_{\ell-1}(I)$. Note how, similarly to eq. (A.16) , eq. (A.32) gives a localization condition.

We can now give a formal definition of *object part* as the subset of the signal I whose complex response, at layer ℓ, is invariant under transformations in the range of the pooling at that layer. This definition is consistent, since the invariance is increasing from layer to layer, therefore allowing bigger and bigger parts. Consequently, for each transformation there exists a layer $\bar{\ell}$ such that any signal subset will be a part at that layer.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

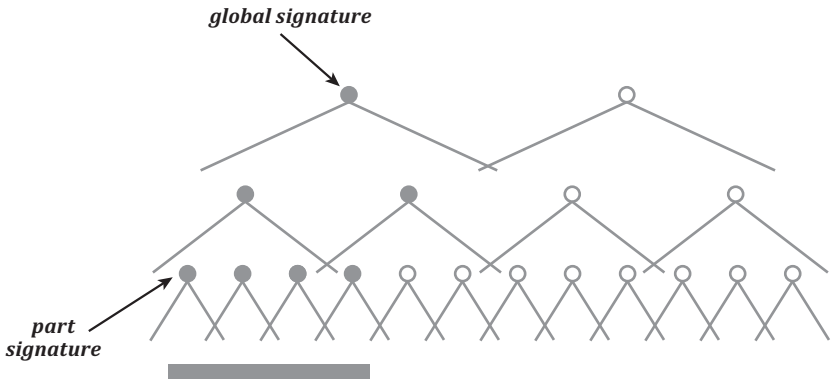

Figure A.3

An image I (rectangle) with a finite support may or may not be fully included in the receptive field of a single complex cell at layer ℓ (and the transformed image may not be included in the pooling range of the complex cell). However, there will be a higher layer such that the support of its neural response is included in the pooling range of a single complex cell.

We can now state the following theorem.

Theorem A.17 (Whole and Parts) Let $I \in \mathcal{X}$ (an image or a subset of it) and μ_ℓ be the complex response at layer ℓ. Let $G_0 \subseteq \cdots \subset G_\ell \subseteq \cdots \subseteq G_L = G$ be a set of nested subsets of the group G. Suppose η is a bijective (positive) function and that the template t and the complex response at each layer has finite support. Then $\forall \bar{g} \in G$, $\mu_\ell(I)$ is invariant for some $\ell = \bar{\ell}$,

$$\mu_m(\bar{g}I) = \mu_m(I), \;\; \exists\, \bar{\ell} \;\text{ s.t. }\; \forall m \geq \bar{\ell}.$$

Proof: The proof follows from the observation that the pooling range over the group is a bigger and bigger subset of G with a growing layer number. In other words, there exists a layer such that the image and its transformations are within the pooling range at that layer (figure A.3). This is clear, since for any $\bar{g} \in G$ the nested sequence

$$G_0 \cap \bar{g}G_0 \subseteq ... \subseteq G_\ell \cap \bar{g}G_\ell \subseteq ... \subseteq G_L \cap \bar{g}G_L = G$$

will include a set $G_{\bar{\ell}} \cap \bar{g}G_{\bar{\ell}}$ such that

$$\left\langle g\mu_{\bar{\ell}-1}(I), t \right\rangle \neq 0 \;\; g \in G_{\bar{\ell}} \cap \bar{g}G_{\bar{\ell}},$$

being $\text{supp}(\left\langle g\mu_{\bar{\ell}-1}(I), t \right\rangle) \subseteq G$.

Theorem A.17 gives a new formal definition of an object part as the part of the visual signal whose representation is invariant at a certain layer of the hierarchical architecture.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A.4 Complex Cells: Wiring and Invariance

We prove the equivalence of signatures obtained by pooling over the projections of template orbits or their principal component. Let $L^2(\mathcal{G}) = \{F : \mathcal{G} \to \mathbb{R} \mid \int |F(g)|^2 dg < \infty\}$, and

$$T_t : \mathcal{X} \to L^2(\mathcal{G}), \qquad (T_t f)(g) = \langle f, T_g t \rangle,$$

where $t \in \mathcal{X}$. It is easy to see that T_t is a linear bounded and compact operator if $\|T_g t\| < \infty$. (In fact, it is easy to see that T is Hilbert-Schmidt, $\mathrm{Tr}(T_t^* T_t) = \int dg \|T_g t\|$.) Denote by $(\sigma_i; u_i, v_i)_i$ the singular system of T_t, where $(u_i)_i$ and $(v_i)_i$ are orthonormal bases for $\mathcal{X}$ and $L^2(\mathcal{G})$, respectively.

For $\sigma : \mathbb{R} \to \mathbb{R}$ measurable, define (complex response)

$$c : \mathcal{X} \to \mathbb{R}, \qquad c(I) = \sum_i \sigma(\langle I, u_i \rangle).$$

If $\sigma(a) = |a|^2$, $a \in \mathbb{R}$, and T_t / b_t is an isometry, where b_t is a constant possibly depending on t (see [131]), then c is invariant. Indeed,

$$c(I) = \|I\|^2 = \frac{1}{b_t^2} \|T_t I\|^2_{L^2(\mathcal{G})} = \frac{1}{b_t^2} \int |\langle I, T_g t \rangle|^2 dg$$

for all $I \in \mathcal{X}$, and $\|T_t I\|^2_{L^2(\mathcal{G})} = \|T_t T_{g'} I\|^2_{L^2(\mathcal{G})}$ for all $g' \in \mathcal{G}$.

If $\mathcal{G}$ is the *affine group* and $\mathcal{X} = L^2(\mathbb{R})$, then under the admissibility condition

$$\int |\langle T_g t, t \rangle|^2 < \infty,$$

it is possible to take $b_t = \sqrt{C_t}$, with $C_t = 2\pi \int \frac{|\hat{t}(\omega)|^2}{\omega} d\omega$, where $\hat{t}$ denotes the Fourier transform of t.

A.5 Mirror-Symmetric Templates Lead to Odd-Even Covariance Eigenfunctions

As explained in section 4.5, in the brain face patch system, specifically in the anterior patch AL, neurons are found to be tuned to even receptive fields. A possible explanation in terms of extracted PCs of a set of rotated-in-depth faces is as follows. Let $R : \mathbb{R}^K \to \mathbb{R}^K$ denote the reflection operator. The eigenfunctions of R are odd and even functions with eigenvalue $-1, 1$, respectively. We want to prove that the covariance matrix associated with the set of templates and their mirror symmetric has odd or even eigenfunctions.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Consider a set that only contains one template t and its reflection $R\tau$ (a unitary transformation). Let Z denote the operator that takes a neural frame T and returns the set consisting of t and its reflection.

$$\begin{aligned} Z &= (\imath, R), \\ Z(t) = \mathbb{T} &= \begin{pmatrix} T \\ Rt \end{pmatrix}, \end{aligned}$$

We want to prove that R commutes with $Z^\intercal Z$, namely, that they have common eigenvectors. We have

$$Z^\intercal Z(Rt) = \begin{pmatrix} Rt + t \\ t + Rt \end{pmatrix} = R(Z^\intercal Z(t)). \tag{A.33}$$

Therefore $R(Z^\intercal Z) = (Z^\intercal Z)R$ (they commute), and $Z^\intercal Z$ and R must have the same eigenfunctions. Since the eigenfunctions of R are even and odd, the principal components of $\mathbb{T}^\intercal \mathbb{T}$ must also be even and odd.

A.6 Gabor-like Shapes from Translation Group

A.6.1 Spectral Properties of Template Transformations Covariance Operator: Cortical Equation

This section focuses on characterizing the spectral properties associated with the covariance of the template transformations. It proposes a cortical equation whose solution provides the eigenfunctions of the covariance of the transformed templates. We consider a layer of 2-D apertures and the covariance of the template transformations associated with each aperture resulting from transformations of images seen through one of these apertures. This leads to an explicit solution for the first layer in the case of translations.

For any fixed t we want to solve the spectral problem associated to the set

$$\mathbb{T}_t = (g_0 t, g_1 t, \ldots, g_{|G|} t)^T,$$

that is, we want to find the eigenvalues λ_i and eigenfunctions ψ_i such that

$$\mathbb{T}_t^* \mathbb{T}_t \psi_i = \lambda_i \psi_i. \tag{A.34}$$

To state the problem precisely we need some definitions. We start with the 1-D problem. We show how to derive an analytical expression of the visual cortex cells tuning based on the following:

- *Observables*. Images, transformed by a locally compact group, seen through an aperture;

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

- *Hebbian learning*. Hebbian-like synapses exist in the visual cortex.

We fix few objects:

- $\mathcal{X}$ space of signals: $L^2(\mathbb{R}, dx)$;
- $\mathcal{T} \subseteq \mathcal{X}$ is the template set.

We solve the eigenproblem associated to the continuous version of eq. (A.34). In this case the basic observable is given by the operator $T : \mathcal{X} \to \mathcal{X}$,

$$(TI)(y) \equiv [t * M_a I](y) = \int dx t(y-x)a(x)I(x), \;\; t \in \mathcal{T},\; a, I \in \mathcal{X}, \tag{A.35}$$

where

$$(M_a I)(x) = a(x)I(x),\; a \in \mathcal{X}.$$

Eq. (A.35) is the mathematical expression of the observable T, that is, a translating template t seen through the function a, which will be called the aperture.

Remark A.18 T is linear (from the properties of the convolution operator) and bounded (from $\|T\| = \|\mathfrak{F}(T)\| = \|t\|\,\|a\|$).

Remark A.19 M_a is a self-adjoint operator. The adjoint operator $T^* : \mathcal{X} \to \mathcal{X}$ is given by

$$\begin{aligned}\left\langle TI, I' \right\rangle &= \int dy\, \bar{I}'(y) \int dx\, t(y-x)a(x)I(x) \\ &= \int dx\, I(x)a(x) \int dy\, t(y-x)\bar{I'}(y) = \left\langle I, T^* I' \right\rangle,\end{aligned}$$

which implies $T^* I = M_a(t^- * I)$, $t^-(x) = t(-x)$. Note that $\|t\| = \left\|t^-\right\| \Rightarrow \|T\| = \|T^*\|$, that is, $\|T^*\|$ is bounded.

Assuming Hebbian learning, we have that the tuning of the cortical cells is given by the solution of the spectral problem of the covariance operator associated to T, $T^*T : \mathcal{X} \to \mathcal{X}$,

$$\begin{aligned}[T^*TI](y) &= M_a[t^- * (t * (M_a I))](y) = M_a[(t^- * t) * (M_a I)](y) \\ &= M_a(t^{\circledast} * (M_a I)) = a(y) \int dx\, a(x) t^{\circledast}(y-x) I(x),\end{aligned}$$

where $t^{\circledast}$ is the autocorrelation of t. This expression can be written as

$$[T^*TI](y) = \int dx\, K(x,y) I(x), \quad K(x,y) = a(x)a(y)t^{\circledast}(y-x).$$

Since the kernel K is Hilbert-Schmidt,

$$Tr(K) = \int dx\, K(x,x) = \int dx\, a^2(x) t^{\circledast}(0) < \infty,$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

we have that

- the eigenfunctions corresponding to distinct eigenvalues are orthogonal;
- the eigenvalues are real and positive;
- there is at least one eigenvalue and one eigenfunction (when K is almost everywhere nonzero), and in general a countable set of eigenfunctions.

Now, we aim to find $\psi_n \in \mathcal{X}$ and $\lambda_n \in \mathbb{R}$ such that

$$a(y) \int dx\, a(x) t^{\circledast}(y-x)\psi_n(x) = \lambda_n \psi_n(y) \tag{A.36}$$

and study their properties. In particular, we find approximate solutions and show that they are Gabor-like shapes. An exact solution in some particular cases can be found in section A.6.2.

A.6.1.1 Square Aperture, Circulants, and Fourier

We start from the simplest discrete case in which we assume periodic boundary conditions on each aperture (one in a layer of receptive fields), resulting in a circulant structure of the collection of transformed templates. Define T as the circulant matrix where each column represents a template t shifted relative to the previous column. This corresponds to assuming that the visual world translates and is seen through a square aperture with periodic boundary conditions. Let us assume in this example that the image is one-dimensional. Thus the image seen through an aperture,

$a(x)$ s.t. $a(x) = 1$ for $0 \leq x \leq A$; $a(x) = 0$ otherwise,

is $t(x-y)a(x)$ when the image is shifted by y. We are led to the following problem. Find the eigenvectors of the symmetric matrix $T^T T$, where T is a circulant matrix (see [132]). If we consider the continuous version of the problem, that is, the eigenvalue problem

$$\int_0^A dx \psi_n(x) t^{\circledast}(y-x)dx = \lambda_n \psi_n(y),$$

with $t^{\circledast}(x)$ being the autocorrelation function associated with t, the solution is $\psi_n(x) = e^{-i2\pi \frac{n}{A}x}$, which is the Fourier basis between 0 and A.

A.6.2 Cortical Equation: Derivation and Solution

Suppose $a(x) = e^{-\alpha x^2}$. Eq (A.36) can be written as

$$\lambda_n \psi_n(y) - e^{-\alpha y^2} \int dx\, e^{-\alpha x^2} t^{\circledast}(y-x)\psi_n(x) = 0,$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

or equivalently,

$$\lambda_n \xi_n(y) - \int dx\, e^{-2\alpha x^2} t^{\circledast}(y-x)\xi_n(x) = 0, \;\; \xi_n(x) = e^{+\alpha x^2}\psi_n(x).$$

Decomposing $t^{\circledast}(x)$ in the Fourier basis,

$$\lambda_n \xi_n(y) - \frac{1}{\sqrt{2\pi}} \int dx\, e^{-2\alpha x^2} \int d\omega\, t^{\circledast}(\omega) e^{i\omega(y-x)} \xi_n(x) = 0.$$

Deriving twice in the y variable and rearranging the expression,

$$\lambda_n \xi_n^{''}(y) + \frac{1}{\sqrt{2\pi}} \int d\omega\, \omega^2 t^{\circledast}(\omega) e^{i\omega y} \int dx\, e^{-2\alpha x^2} \psi_n(x) e^{-i\omega x} = 0. \qquad \text{(A.37)}$$

This expression is equivalent to the original eigenproblem in eq. (A.36) adding the conditions (e.g., $\xi_n(0) = c$, $\xi_n^{'}(0) = c^{'}$). In fact, given the derivative of a function, the original function is uniquely determined, having its value on a point. Therefore the function $\xi_n^{'}(y)$ uniquely determines $\xi_n(y)$ if we have, for instance, $\xi_n(0) = c$, and the function $\xi_n^{''}(y)$ uniquely determines $\xi_n^{'}(y)$ if we have, for instance, $\xi_n^{'}(0) = c^{'}$.

Using the Fourier operator $\mathfrak{F}$, we can rewrite (A.37) as

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\mathfrak{F}^{-1}\Big(\omega^2 t^{\circledast}(\omega)\mathfrak{F}\big(e^{-2\alpha x^2}\psi_n(x)\big)\Big) = 0.$$

By the convolution theorem,

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\mathfrak{F}^{-1}\big(\omega^2 t^{\circledast}(\omega)\big) * \big(e^{-2\alpha x^2}\xi_n(x)\big) = 0.$$

Expanding $\omega^2 t^{\circledast}(\omega)$ in a Taylor series and recalling that $\mathfrak{F}^{-1}(\omega^m) = i^m \delta^m(x)$, we are led to the differential equation

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\big(c_0\delta + ic_1\delta^{'} + \cdots\big) * \big(e^{-2\alpha x^2}\xi_n(x)\big) = 0, \qquad \text{(A.38)}$$

where c_i are the coefficients of the Taylor series expansion of $\omega^2 t^{\circledast}(\omega)$.

A.6.2.1 $1/\omega$ Spectrum

In the case of the average natural images spectrum

$$t(\omega) = \frac{1}{\omega},$$

the differential eq. (A.38) assumes the particularly simple form

$$\lambda_n \xi_n^{''}(y) + \sqrt{2\pi}\, e^{-2\alpha y^2}\psi_n(y) = 0.$$

In the harmonic approximation

$$\xi_n^{''}(y) + \frac{\sqrt{2\pi}}{\lambda_n}(1 - 2\alpha y^2)\xi_n(y) = 0, \qquad \text{(A.39)}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

which is of the form of a Weber differential equation (or 1-D Schrödinger equation for the harmonic oscillator). The explicit real solution is given by

$$\xi_n(y) = D\Big(-\frac{1}{2} + \frac{\pi^{\frac{1}{4}}}{2^{\frac{5}{4}}\sqrt{\alpha\lambda_n}}, \frac{2^{\frac{7}{8}}\alpha^{\frac{1}{4}}\pi^{\frac{1}{8}}}{\lambda_n{}^{\frac{1}{4}}}\Big),$$

where $D(\eta, y)$ is the parabolic cylinder function. It can be proved [134, p.139] that the solutions have two different behaviors, exponentially increasing or exponentially decreasing, and that we have exponentially decreasing solutions if and only if the following quantization condition holds:

$$-\frac{1}{2} + \frac{\pi^{\frac{1}{4}}}{2^{\frac{5}{4}}\sqrt{\alpha\lambda_n}} = n,\ n = 0, 1, \ldots.$$

Therefore, recalling that $\alpha = 1/\sigma_\alpha^2$, we obtain the spectrum quantization condition

$$\lambda_n = \sqrt{\frac{\pi}{2}}\frac{\sigma_\alpha^2}{(2n+1)^2},\ n = 0, 1, \ldots. \tag{A.40}$$

Further, using the identity (true if $n \in \mathbb{N}$)

$$D(n, y) = 2^{-\frac{n}{2}} e^{-\frac{y^2}{4}} H_n\Big(\frac{y}{\sqrt{2}}\Big),$$

where $H_n(y)$ are Hermite polynomials, we have

$$\xi_n(y) = 2^{-\frac{n}{2}} e^{-\frac{2n+1}{\sigma_\alpha^2}y^2} H_n\Big(\frac{\sqrt{2(2n+1)}}{\sigma_\alpha}y\Big),$$

or (see figure A.4)

$$\psi_n(y) = 2^{-\frac{n}{2}} e^{-\frac{2(n+1)}{\sigma_\alpha^2}y^2} H_n\Big(\frac{\sqrt{2(2n+1)}}{\sigma_\alpha}y\Big). \tag{A.41}$$

Remark A.20 The solution in eq. (A.41) is also an approximate solution for any template spectrum such that

$$\omega t(\omega) = const + O(\omega).$$

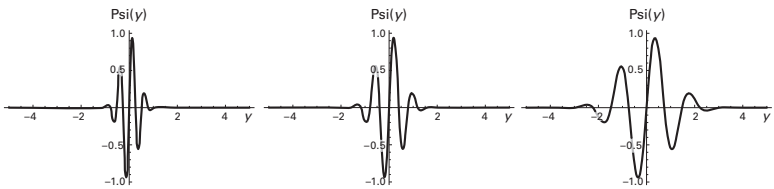

Figure A.4

Wavelet-like behavior of the solution for $\sigma_\alpha/\sqrt{\lambda} = C$. $C = 10,\ \lambda = 5, 10, 20$.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A.6.2.2 Extension to the Two-Dimensional Problem

Suppose the template t is moving only in the x direction. The spectral problem in 2-D is

$$\lambda_n \xi_n(x,y) - \int dx\, e^{-2\alpha(\eta^2+\gamma^2)} t^{\circledast}(x-\eta,y)\xi_n(\eta,\gamma) = 0,$$
$$\xi_n(x,y) = e^{+\alpha(x^2+y^2)}\psi_n(x,y).$$

Applying the operator ∇^2 and proceeding as in the 1-D case, we obtain

$$\lambda_n \nabla^2 \xi_n(x,y) + 2\pi \mathfrak{F}^{-1}\big((\omega_\eta^2+\omega_\gamma^2)t^{\circledast}(\omega_\eta,\omega_\gamma)\big) * \big(e^{-2\alpha(\eta^2+\gamma^2)}\xi_n(\eta,\gamma)\big)_{|_{y=0}} = 0.$$

If we suppose the spectrum is of the form

$$t(\omega_x,\omega_y) = \frac{1}{\sqrt{\omega_x^2+\omega_y^2}},$$

we are left with the differential equation

$$\nabla^2\xi(x,y) + \frac{2\pi}{\lambda_n} e^{-2\alpha x^2}\xi_n(x,0) = 0. \tag{A.42}$$

In the hypothesis $\xi_n(x,y) = \xi_{1,n}(x)\xi_{2,n}(y)$ we can immediately see, plugging the solution into eq. (A.42), that $\xi_n(x,y)$ is y-independent. Therefore the solution in the 2-D case is

$$\psi_n(x,y) = 2^{-\frac{n}{2}} e^{-\frac{y^2+2(n+1)x^2}{\sigma_\alpha^2}} H_n\Big(\frac{\sqrt{2(2n-1)}}{\sigma_\alpha}x\Big), \tag{A.43}$$

which gives

$$\frac{\sigma_y}{\sigma_x} = \sqrt{2(n+1)}.$$

A.7 Gabor-like Wavelets

A.7.1 Derivation from an Invariance Argument for 1-D Affine Group

The equality in eq. (A.19) in general is not true. However, this is clearly the case if we consider the group of transformations to be dilations and translations (see figure A.5a). We discuss this case in some detail.

Assume that $G_0 = [0,a]$. Let

$$S = \{\langle T_x I, t\rangle \ : \ x \in [0,a]\}, \ \ S_c = \{\langle T_x I, t\rangle \ : \ x \in [c,a+c]\} \tag{A.44}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

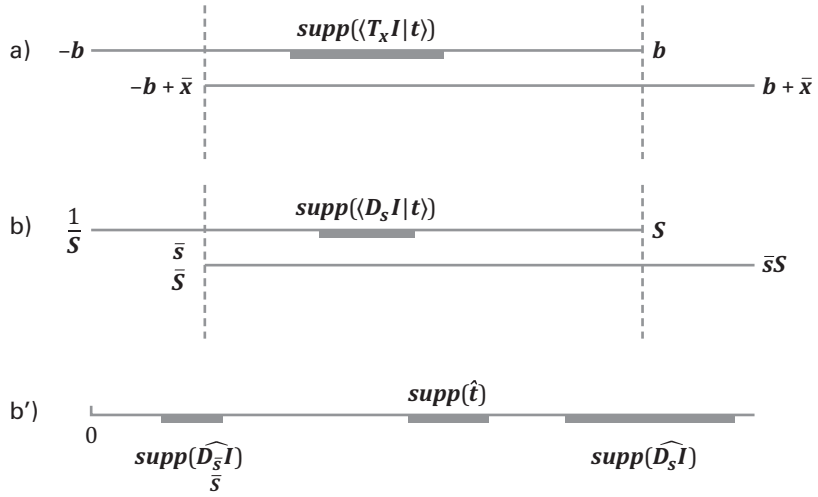

Figure A.5

If the support of the dot product between the image and the template is contained in the intersection between the pooling range and the group-translated (*a*) or dilated (*b*) pooling range the signature is invariant. In frequency condition (*b*) becomes (*b*′): When the Fourier supports of the dilated image and the template do not intersect, their dot product is zero.

for a given c, where T_x is a unitary representation of the translation operator. We can view S, S_c as sets of simple responses to a given template through two receptive fields. Let $S_0 = \{\langle T_x I, t \rangle \ : \ x \in [0, a + c]\}$, so that $S, S_c \subset S_0$ for all c. We assume S_0, S, S_c to be closed intervals for all c. Then recall that a bijective function (in this case η_h) is strictly monotonic on any closed interval, so the difference of integrals in eq. (A.19) is zero if and only if $S = S_c$. Since we are interested in considering all the values of c up to some maximum C, we can consider the condition

$$\langle T_x I, t \rangle = \langle T_x T_a I, t \rangle \, , \ \forall \, x \in [0, c], \ c \in [0, C]. \tag{A.45}$$

This condition can be satisfied in two cases: (1) both dot products are zero, which is the localization condition, or (2) $T_a I = I$ (or equivalently, $T_a t = t$), where the image or template is periodic. A similar reasoning applies to the case of scale transformations.

In the next section we show how localization conditions for scale and translation transformations imply a specific form of the templates.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A.7.2 Wavelet Templates from Best Invariance and Heisenberg Principle

In this section we identify G_0 with subsets of the affine group. In particular, we study separately the case of scale and translations (in 1-D for simplicity). It is helpful to assume that all images I and templates t are strictly contained in the range of translation or scale pooling, P, since image components outside it are not measured. We consider images I *restricted to* P. For translation this means that the support of I is contained in P. For scaling, since $g_s I = I(sx)$ and $\widehat{I(sx)} = (1/s)\hat{I}(\omega/s)$ (where $\hat{\cdot}$ indicate the Fourier transform), assuming a scale pooling range of $[s_m, s_M]$, implies a range $[\omega_m^I, \omega_M^I]$, $[\omega_m^t, \omega_M^t]$ (m and M indicate maximum and minimum) of spatial frequencies for the maximum support of I and t. Because of theorem A.8 *invariance to translation requires spatial localization of images and templates*, and *invariance to scale requires bandpass properties of images and templates*. Thus images and templates are assumed to be localized from the outset in either space or frequency. The following corollaries show that a stricter localization condition is needed for invariance and that this condition determines the form of the template. Note that in our framework images and templates are bandpass because of being zero-mean. In addition, neural images, which are input to the hierarchical architecture, are spatially bandpass because of retinal processing.

We now state the result of theorem A.8 for one-dimensional signals under the translation group and separately under the dilation group.

Let $I, t \in L^2(\mathbb{R})$, $(\mathbb{R}, +)$ be the one-dimensional locally compact group of translations and $T_x : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ be a unitary representation of the translation operator. For instance, $G_0 = [-b, b]$, $b > 0$, and suppose $\text{supp}(t) \subseteq \text{supp}(I) \subseteq [-b, b]$. Further suppose $\text{supp}(\langle T_x I, t\rangle) \subseteq [-b, b]$. Then we have the following.

Corollary A.21 *Localization in the spatial domain is necessary and sufficient for translation invariance.*

For any fixed $t, I \in \mathcal{X}$, we have

$$\mu_h^k(I) = \mu_h^k(T_x I),\ \forall x \in [0, \bar{x}] \ \Leftrightarrow\ \langle T_x I, t\rangle \neq 0,\ \forall x \in [-b + \bar{x}, b] \tag{A.46}$$

with $\bar{x} > 0$.

Similarly, let $G = (\mathbb{R}^+, \cdot)$ be the one-dimensional locally compact group of dilations, and denote by $D_s : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ a unitary representation of the dilation operator. Let $G_0 = [1/S, S]$, $S > 1$ and suppose $\text{supp}(\langle D_s I, t\rangle) \subseteq [1/S, S]$. Then eq. (A.8) gives the following.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Corollary A.22 *Localization in the spatial frequency domain is necessary and sufficient for scale invariance.*

For any fixed $t, I \in \mathcal{X}$, we have

$$\mu_h^k(I) = \mu_h^k(D_s I), s \in [1, \bar{s}] \;\; \Leftrightarrow \;\; \langle D_s I, t \rangle \neq 0, \; \forall s \in [\frac{\bar{s}}{S}, S] \tag{A.47}$$

with $S > 1$.

Localization conditions of the support of the dot product for translation and scale are depicted in figure A.5.

As shown by lemma A.23, eqs. (A.46) and (A.47) give interesting conditions on the supports of t and its Fourier transform $\hat{t}$. For translation the corollary is equivalent to zero overlap of the compact supports of I and t. In particular, using theorem A.5, for $I = t$ the maximal invariance implies the following localization conditions on t,

$$\langle gt, t \rangle = 0 \;\; g \notin G_L \subseteq G, \tag{A.48}$$

which we call self-localization. For 1-D translations it has the simple form $\langle T_x t, t \rangle = 0 \; |x| > a, \; a > 0$.

For scaling we consider the support of the Fourier transforms of I and t. The Parseval theorem allows rewriting the dot product $\langle D_s I, t \rangle$, which is in $L^2(\mathbb{R}^2)$, as $\langle \widehat{D_s I}, \hat{t} \rangle$ in the Fourier domain.

In the following we suppose that the supports of $\hat{t}$ and $\hat{I}$ are, respectively, $[\omega_m^t, \omega_M^t]$ and $[\omega_m^I, \omega_M^I]$, where $\omega_m^{t,I}$ could be very close to zero (images and templates are supposed to be zero-mean) but usually are larger than zero. Note that the effect of scaling I with (typically $s = 2^j$ with $j \leq 0$) is to change the support to $\text{supp}(\widehat{D_s I}) = s(\text{supp}(\hat{I}))$. This change of the support of $\hat{I}$ in the dot product $\langle \widehat{D_s I}, \hat{t} \rangle$ gives nontrivial conditions on the intersection with the support of $\hat{t}$ and therefore on the localization with respect to the scale invariance. We have the following lemma.

Lemma A.23 Invariance to translation in the range $[0, \bar{x}], \; \bar{x} > 0$ is equivalent to the localization condition of t in space

$$\text{supp}(t) \subseteq [-b - \bar{x}, b] - \text{supp}(I), \; I \in \mathcal{X}. \tag{A.49}$$

Separately, invariance to dilation in the range $[1, \bar{s}], \; \bar{s} > 1$ is equivalent to the localization condition of $\hat{t}$ in frequency ω

$$\begin{aligned} &\text{supp}(\hat{t}) \subseteq [-\omega_t - \Delta_t^*, -\omega_t + \Delta_t^*] \cup [\omega_t - \Delta_t^*, \omega_t + \Delta_t^*] \\ &\Delta_t^* = S\omega_m^I - \omega_M^I \frac{\bar{s}}{S}, \omega_t = \frac{\omega_M^t - \omega_m^t}{2}. \end{aligned} \tag{A.50}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Proof: To prove that $\text{supp}(t) \subseteq [-b + \bar{x}, b] - \text{supp}(I)$, note that eq. (A.46) implies that $\text{supp}(\langle T_x I, t\rangle) \subseteq [-b + \bar{x}, b]$ (see figure A.5*a*). With $\text{supp}(\langle T_x I, t\rangle) = \text{supp}(I * t) \subseteq \text{supp}(I) + \text{supp}(t)$, we have $\text{supp}(t) \subseteq [-b - \bar{x}, b] - \text{supp}(I)$. To prove the condition in eq. (A.50),note that eq. (A.47) is equivalent in the Fourier domain to

$$\langle D_s I, t\rangle = \langle \widehat{D_s I}, \hat{t}\rangle = \frac{1}{s}\int d\omega\, \hat{I}\big(\frac{\omega}{s}\big)\hat{t}(\omega) \neq 0 \;\; \forall s \in [\frac{\bar{s}}{S}, S]. \tag{A.51}$$

This is depicted in figure A.5b' for S large enough. In this case we can suppose the support of $\widehat{D_{\bar{s}/S} I}$ to be on an interval to the left of $\text{supp}(\hat{t})$, and $\widehat{D_S I}$ to the right. The condition $\text{supp}(\langle \widehat{D_s I}, \hat{t}\rangle) \subseteq [\bar{s}/S, S]$ is in this case equivalent to

$$\omega_M^I \frac{\bar{s}}{S} < \omega_m^t, \;\; \omega_M^t < \omega_m^I S, \tag{A.52}$$

which gives

$$\Delta_t^* = \text{Max}(\Delta_t) \equiv \text{Max}\Big(\frac{\omega_M^t - \omega_m^t}{2}\Big) = S\omega_m^I - \omega_M^I \frac{\bar{s}}{S}. \tag{A.53}$$

and therefore eq. (A.50).

Note that for some $s \in [\bar{s}/S, S]$ the condition that the Fourier supports are disjoint is only sufficient and not necessary for the dot product to be zero, since cancellations can occur. However, to have $\langle \widehat{D_s I}, \hat{t}\rangle = 0$ on a continuous interval of scales (except some pathological examples of the function I) implies disjointness of the supports, since $\hat{I}(\omega/s) \neq \hat{I}(\omega/s')$, $s \neq s'$, unless I has constant spectrum in the interval $[\bar{s}/S, S]$. Similar reasoning is valid for the translation case.

The preceding results lead to a statement connecting invariance with localization of the templates.

Theorem A.24 Maximum translation invariance implies a template with minimum support in the space domain (x). Maximum scale invariance implies a template with minimum support in the Fourier domain (ω).

Proof: We first illustrate the statement of the theorem with a simple example. In the case of translations, suppose, for instance, $\text{supp}(I) = [-b, b']$, $\text{supp}(t) = [-a, a]$, $a \leq b' \leq b$. Then eq. (A.49) reads

$$[-a, a] \subseteq [-b + \bar{x} + b', b - b'],$$

which gives the condition $-a \geq -b + b' + \bar{x}$, that is, $\bar{x}^{\max} = b - b' - a$. Thus for any fixed b, b', the smaller the template support $2a$ in space, the greater is translation invariance. Similarly, in the case of dilations, increasing the range

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

of invariance $[1, \bar{s}]$, $\bar{s} > 1$, implies a decrease in the support of $\hat{t}$, as shown by eq. (A.53). In fact, noting that $|\text{supp}(\hat{t})| = 2\Delta_t$, we have

$$\frac{d|supp(\hat{t})|}{d\bar{s}} = -\frac{2\omega_M^I}{S} < 0,$$

that is, the measure $|\cdot|$ of the support of $\hat{t}$ is a decreasing function with respect to the measure of the invariance range $[1, \bar{s}]$.

Because of the assumption of maximum possible support of all I being finite, there is always localization for any choice of I and t under spatial shift. Of course, if the localization support is larger than the pooling range there is no invariance. For a complex cell with pooling range, $[-b, b]$ in space, only templates with self-localization smaller than the pooling range make sense. An extreme case of self-localization is $t(x) = \delta(x)$, corresponding to maximum localization of tuning of the simple cells.

A.8 Transformations with Lie Group Local Structure: HW Module

In this section we generalize the reasoning of chapter 1 for groups to transformations that generate manifolds with a high degree of smoothness.

Let $I \in \mathcal{X}$. Let $s : \mathcal{X} \times \mathbb{R}^N \to \mathcal{X}$, be a C^∞ transformation depending on $\Theta = (\theta_1, \cdots, \theta_N)$ parameters. For any fixed $I \in \mathcal{X}$, the set $M = (s(I, \Theta),\ \Theta \in \mathbb{R}^N)$ describes a differentiable manifold.

If we expand the transformation around, for instance, $\vec{0}$, we have

$$s(I, \Theta) = s(I, \vec{0}) + \sum_{i=1}^{N} \frac{\partial s(I, \Theta)}{\partial \theta_i}\theta_i + o(\|\Theta\|^2) = I + \sum_{i=1}^{N} \theta_i L_{\theta_i}(I) + o(\|\Theta\|^2), \tag{A.54}$$

where L_{θ_i} are the infinitesimal generators of the transformation in the i^{th} direction. The associated transformation can be expressed by exponentiation as

$$g(\Theta) = \exp(\theta_1 L_{\theta_1} + \theta_2 L_{\theta_2} + \cdots + \theta_N L_{\theta_N}).$$

Consider now the signature

$$\mu^{h,k}(I) = \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle s(I, \Theta), t^k \right\rangle). \tag{A.55}$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

If the support of the template is localized enough with respect to that of the image, we have (suppose for simplicity the generators L_{θ_i} are self-adjoint)

$$\mu^{h,k}(I) = \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle s(I,\Theta), t^k \right\rangle \approx \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle g(\Theta)I, t^k \right\rangle$$
$$= \int_{\mathbb{R}^N} d\Theta\ \eta_h(\left\langle I, g(\Theta)t^k \right\rangle.$$

If the representation of the group associated to the operators $g(\Theta)$ is square integrable (a continuous irreducible unitary representation $\pi(g) = g(\Theta)$ of a locally compact group G in a Hilbert space H such that for some nonzero vector ψ the function, $g \to \langle g(\Theta)\psi, \psi \rangle$, is square integrable with respect to the Haar measure on G), then $g(\Theta)t^k$ is the associated wavelet frame. The signature calculated in this way is invariant and selective (see chapter 1).

A.9 Factorization of Invariances

Initially we thought that a signature invariant to a group of transformations could be obtained by factorizing in successive layers the computation of signatures invariant to a subgroup of the transformations (e.g., the subgroup of translations of the affine group) and then adding the invariance with respect to another subgroup (e.g., rotations). In the following we show that while factorization of invariance ranges is possible in a hierarchical architecture (theorem A.17), in general the factorization in successive layers for instances of invariance to translation followed by invariance to rotation (by subgroups) is actually impossible. The mathematical reason is that subgroup averages destroy the covariance structure, factoring out the transformation.

We begin by considering invariance with respect to 2-D translations. We show how invariance can be achieved by first computing an invariant signature with respect to the x-translation and the y-translation, and discuss how a similar strategy can be used with more general groups (and subgroups). We then show that this factorization cannot be used to learn separately x and y invariance through independent templates.

Consider an architecture composed of two layers. Each layer has an associated set of templates, $t^1, \ldots, t^K \in \mathcal{X}$ and $\bar{t}^1, \ldots, \bar{t}^K \in \mathcal{X}$ (we assume for simplicity the number of templates is the same). We also assume we have just one nonlinearity per layer and that the set of translations has been suitably discretized. We denote by T_x, T_y, $x, y \in \mathbb{R}$, the action of the translations on any $I \in \mathcal{X}$.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

The first layer defines an invariant signature $\mu : \mathcal{X} \to \mathbb{R}^K$ as

$$\mu_1^k(I) = \sum_{x' \in \mathbb{R}} \eta(\left\langle T_{x'} I, t^k \right\rangle), \quad k = 1, ..., K, \tag{A.56}$$

where $\langle T_{x'} I, t^k \rangle = \left\langle I, T_{x'}^{-1} t^k \right\rangle$, since the representation of the translation is unitary. This means that we can obtain the signature by looking at transforming images or transforming templates.

The signature at second layer is $\nu : \mathcal{X} \to \mathbb{R}^K$,

$$\mu_2^l(I) = \sum_{y' \in \mathbb{R}} \eta(\left\langle \mu_1(T_{y'} I), s^l \right\rangle), \quad l = 1, ..., K, \tag{A.57}$$

where the set of templates $s^1, \ldots, s^K$ can be thought of as the signatures of a set of templates with respect to the first layer, that is, $s^i = \mu_1(\bar{t}^i)$ for $\bar{t}^i \in \mathcal{X}$, $i = 1, \ldots, K$, with $s^i \in \mathbb{R}^K$.

Indeed, we can show that ν is invariant to 2-D translations $\mu_2(I) = \mu_2(T_x T_y I)$. We note that $\mu(T_x T_y I) = \mu(T_y T_x I) = \mu(T_y I)$, since μ is defined by a group integration with respect to the x-translation and the x and y translation operators commute. Then

$$\mu_2^l(T_x T_y I) = \sum_{y' \in \mathbb{R}} \eta(\left\langle \mu_1(T_{y'} T_y I), s^l \right\rangle) = \mu_2^l(T_y I), \quad l = 1, \ldots, K, \tag{A.58}$$

and finally $\mu_2^l(T_x T_y I) = \mu_2^l(T_y I) = \mu_2(I)$, since μ_2 is defined by a group integration with respect to the y-translation.

An inspection of the reasoning shows that a similar factorization holds for many transformations beyond translations. Indeed, we have the following result.

Lemma A.25 (Factorization of Invariances) Let G, R be two locally compact groups with Haar measures dg, dr, respectively. Let $\mu : \mathcal{X} \to \mathbb{R}^K$ and $\nu : \mathcal{X} \to \mathbb{R}^L$ be defined by

$$\mu_1^k(I) = \int dg \eta(\left\langle gI, t^k \right\rangle), \qquad \mu_2^l(I) = \int dr \eta(\left\langle \mu_1(rI), s^l \right\rangle); \tag{A.59}$$

then $\mu_2(grI) = \mu_2(I)$.

Remark A.26 We briefly comment on the order in which transformations need to be applied to an image to still have invariance. Clearly, factorization of invariances happens in the architecture with an order given by the group integration at each layer, so in general it might not be true that $\mu_2(rgI) = \mu_2(I)$. However, invariance clearly still holds if the groups actions commute, $rg = gr$. Moreover it also holds under the weaker requirement that for all

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

$(g,r) \in G \times R$ there exist $(g',r') \in G \times R$ such that $rg = g'r'$. The latter property holds, for example, if we take G, R to be 2-D translations and rotations, respectively. This is because the semidirect product of the Abelian group $\mathbb{R}^2$ (which describes translations) and the group $SO(2)$ of orthogonal 22 matrices (which describes rotations and reflections that keep the origin fixed) is isomorphic to the Euclidean group. The group of translations is a normal subgroup of the Euclidean group, and the Euclidean group is a semidirect product of the translation group and $SO(2)$.

A more interesting result would be if μ_2 of eq. (A.57) were covariant in the following sense: $\langle \mu_1(g_{y'}I), s^k \rangle) = \left\langle \mu_1(I), \mu_1(g_{y'}^{-1}\bar{t}^i) \right\rangle$, which we rewrite as

$$\left\langle \mu_1(rI), \mu_1(r\bar{t}) \right\rangle = \left\langle \mu_1(I), \mu_1(\bar{t}) \right\rangle, \tag{A.60}$$

where r is the group element corresponding to $g_{y'}$. If this were true, then one invariance could be learned via the templates t^k and another could be learned in the next layer via the templates s^i. This covariance property, however, cannot be expected for general η.

To prove the statement, we first sketch the linear situation: η is the identity function. We also assume that the set of t^k is an orthonormal basis. In this case,

$$\begin{aligned} \left\langle \mu_1(rI), \mu_1(r\bar{t}) \right\rangle &= \sum_k \int dg \left\langle grI, t^k \right\rangle \int dg' \left\langle g'r\bar{t}, t^k \right\rangle \\ &= \int dg \int dg' \sum_k \left\langle grI, t^k \right\rangle \left\langle g'r\bar{t}, t^k \right\rangle, \end{aligned}$$

and if the transformations g and r do commute, we have

$$\begin{aligned} \left\langle \mu_1(rI), \mu_1(r\bar{t}) \right\rangle &= \int dg \int dg' \left\langle grI, g'r\bar{t} \right\rangle \\ &= \int dg \int dg' \left\langle gI, g'r^{-1}r\bar{t} \right\rangle \\ &= \int dg \int dg' \left\langle gI, g'\bar{t} \right\rangle \\ &= \left\langle \mu_1(I), \mu_1(\bar{t}) \right\rangle. \end{aligned} \tag{A.61}$$

Note that if the transformations g and r do not commute, the result in eq. (A.61) does not hold. Even a weaker result, $\left\langle \mu(rI), \mu(r\bar{t}) \right\rangle = \left\langle \mu(I), \mu(r'\bar{t}) \right\rangle$ for some r', does not hold. In fact, using remark A.26, we have that for each fixed g, g', r there exists $\tilde{g}, \hat{g}, r'(g,g')$ such that

$$\left\langle grI, g'r\bar{t} \right\rangle = \left\langle \tilde{g}I, \hat{g}r'(g,g')\bar{t} \right\rangle.$$

However, r' is different for any choice of g, g'. Therefore we can neither obtain the same result as in the commutative case nor have a weaker form of it,

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

$\langle \mu_1(rI), \mu_1(r\bar{t}) \rangle = \langle \mu_1(I), \mu_1(r'\bar{t}) \rangle$ for some r'. Another important case when eq. (A.60) does not hold is when η is a nonlinear function. we now sketch another case of practical interest in which $\eta(x) = x^2$. We make the same simplifying assumptions. Then

$$\begin{aligned} \langle \mu_1(rI), \mu_1(r\bar{t}) \rangle &= \sum_k (\int dg \left\langle grI, t^k \right\rangle)^2 (\int dg' \left\langle g' r\bar{t}, t^k \right\rangle)^2 \\ &= \int dg \int dg' \sum_k (\left\langle grI, t^k \right\rangle)^2 (\left\langle g' r\bar{t}, t^k \right\rangle)^2. \end{aligned}$$

It can be easily proved that this expression does not satisfy eq. (A.60) even in the case of commutative groups. Thus *for the specific network* here, factorization holds only for functions η that are linear (or equivalently, only for the first moment of the distribution) and only for commutative groups.

In this section we analyzed in detail in which sense a hierarchical architecture achieves larger and larger invariance with growing layer numbers. We found that it is impossible to achieve invariance with respect to different subgroups at different layers and that only invariance per ranges is possible.

A.10 Invariant Representations and Their Cost in Number of Orbit Elements

In the model described so far, implementing an invariant (and discriminative) representation has a *memory cost* in terms of number of templates and their transformations that we need to store to calculate an invariant signature. Here we calculate this cost for two different architectures, hierarchical and shallow (figure A.6), and show how a hierarchical architecture can calculate an invariant signature with far fewer transformed templates than a shallow architecture can.

The intuition is the following. Fewer template transformations are needed to compute a certain invariance at layer ℓ in the case of an invariant hierarchical representation because the dimensionality of the space of possible images is reduced owing to the inherited invariance from the layers below.

More precisely, consider the signature

$$\mu_\ell^{h,k}(I)(g) = \int_{gG_\ell} d\bar{g}\ \eta_h(\left\langle \mu_{\ell-1}(I), \mu_{\ell-1}(\bar{g}t^k) \right\rangle), \quad k = 1, \cdots, K. \tag{A.62}$$

As proved previously, $\mu_\ell^{h,k}(I)$ is an invariant and discriminative vector. The amount of invariance (which we call $\bar{G}_\ell$), as proved in theorem A.8, depends on the extension of the support of the dot product $\langle \mu_{\ell-1}(I), \mu_{\ell-1}(\bar{g}t^k) \rangle$ with

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

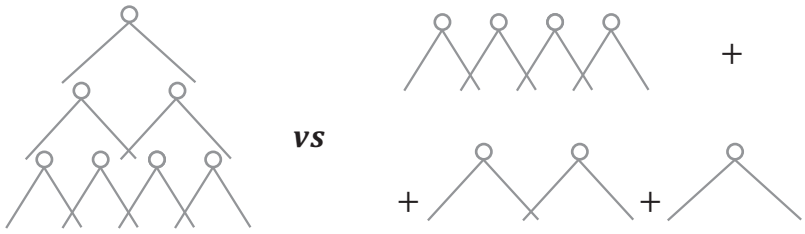

Figure A.6
Hierarchical architecture versus multiple one-layer architectures.

respect to the pooling range G_ℓ. As is clear from eq. A.62, building invariant representations has a cost in terms of the template transformations needed to compute the invariance and a cost in terms of the number of different templates (projections, K) if we want the representation to be discriminative. In the following we consider the cost of hierarchical invariant representations versus multiple one-layer representations achieving the same invariances for whole images and their parts and with the same discriminative power (see figure A.6).

Suppose, instead of a single-layer $\bar{G}$-invariant representation, we construct a sequence of increasingly invariant representations (at different layers of a hierarchy) into the tower of transformation subsets (figure A.7):

$$\bar{G}_1 \subset \bar{G}_2 \subset \cdots \subset \bar{G}_L = \bar{G} \subset G.$$

This construction not only achieves $\bar{G}$ invariance at the top layer for the whole image but also local and *independent* $\bar{G}_\ell$ invariances for (bigger and bigger) parts of the image. If we wanted to achieve the same kind of invariances with a single-layer architecture, we would need to have L one-layer architectures that implement separately the $\bar{G}_\ell$, $\ell = 1, \cdots, L$, invariances for bigger and bigger parts of the image. In the following we mathematically show that the cost (in terms of transformed templates) is far less with a hierarchical architecture than with multiple one-layer architectures. In the next section we consider the discriminative properties of the two architectures and prove that the number of templates to achieve the same discriminability is the same in the two models.

A.10.1 Discriminability Cost for Hierarchical versus Single-Layer Architecture

For simplicity we study the case of a non-Abelian locally compact discrete group of transformations G and its associated wavelet transform (e.g., $g_{j,k} \in G$,

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

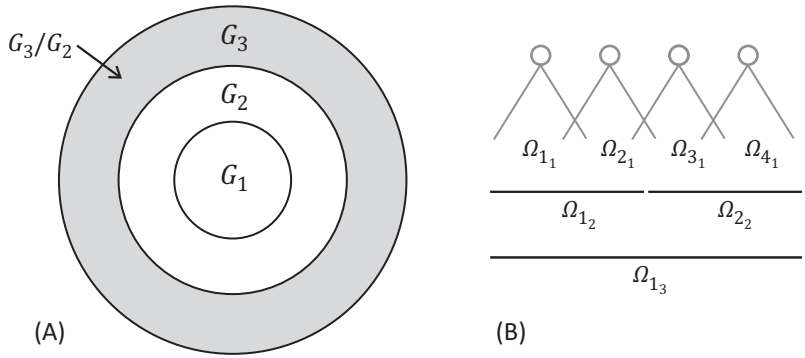

Figure A.7

(A) Tower of subsets of transformations. (B) Effective (at the image level) partitions of the image support (corresponding to different image parts) at different layers of a hierarchical architecture.

$j, k \in \mathbb{Z}$, such that $g_{j,k}f(x) = 2^{j/2}f(2^j x - k)$, $f \in L^2(\mathbb{R})$). It is well known that the number of random projections needed to distinguish among n images with ϵ precision is proportional to $\ln(n)$ and that this number does not depend on the image dimensionality by the Johnson-Lindenstrauss lemma. In section A.2 we proved that if n is the cardinality of a set of different images (none is the transform of another; G-invariant) that must be we have to distinguished with probability $1 - \delta^2$ and precision ϵ, then an estimated probability distribution obtained with

$$\epsilon(K) = \sqrt{\frac{2}{Kc} \ln\left(\frac{n}{\delta}\right)}$$

is sufficient.

Translated in the context of a hierarchical construction this means that the set of images at each module at layer ℓ that must be distinguished are all possible $\bar{G}_\ell$-invariant images, n_ℓ. This is difficult to calculate, but we can prove that the number of equivalence classes in the case of parallel architecture and hierarchical architecture is the same at each layer and part. The only difference is that in the case of a hierarchy the space of the signals at layer ℓ is already $\bar{G}_{\ell-1}$ invariant.

In general, n_ℓ depends, at each module of layer ℓ, on the extension of the support of $\langle \mu_\ell(I), \mu_\ell(gt^k) \rangle$ with respect to the pooling range G_ℓ. In particular, the invariance range is given by

$$\bar{G}_\ell = \{\hat{g} \in G \mid \text{supp}(\left\langle \mu_\ell(\hat{g}I), \mu_\ell(gt^k) \right\rangle \subseteq G_\ell\}.$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

However, note that

$$\text{supp}_{g\in G_\ell}(\langle\mu_\ell(I),\mu_\ell(gt^k)\rangle) = \text{supp}_{g\in G_\ell}(\langle I, gt^k\rangle).$$

Thus we have that $\bar{G}_\ell = G_\ell$ and therefore that the number of G_ℓ ($\bar{G}_\ell$)-invariant images is the same for the two architectures, namely, $k_\ell = \bar{k}_\ell$, with the pooling range, G_ℓ, the same. In other words, we proved that the number of templates per layer and per module to achieve discriminability among the $\bar{G}_\ell$-invariant images is the same in the case of hierarchical (k_ℓ) and parallel ($\bar{k}_\ell$) architectures.

A.10.2 Invariance Cost for Hierarchical versus Single-Layer Architecture

For simplicity, we again consider a model where the transformations are those of the discrete non-Abelian locally compact group of transformations G associated to the discrete wavelet transform in $L^2(\mathbb{R})$.

Theorem A.27 Let $I \in \mathcal{X}^\Omega$ and $\Omega_{i_\ell} \subset \Omega$, $i_\ell = 1,\cdots,M_\ell$, $\ell = 1,\cdots,L$ (see figure A.7, B). Let Ω_{i_ℓ} be such that $\bigcup_{i_\ell=1}^{M_\ell} \Omega_{i_\ell} = \Omega$, $\bigcap_{i_\ell=1}^{M_\ell} \Omega_{i_\ell} = \emptyset$, $\forall\, \ell = 1,\cdots,L$, and suppose the sets are all of the same size per fixed layer. Let $\bar{G}_1 \subset \bar{G}_2 \subset \cdots \subset \bar{G}_L = \bar{G} \subset G$ be a tower of proper subsets of G. Let $c(1)$, $c(L)$ be the costs respectively, of a multiple one-layer and of a hierarchy composed of L layers architecture for achieving the same $\bar{G}_\ell$-invariances (for image parts with support in Ω_{i_ℓ}). We have

$$r = \frac{c(L)}{c(1)} = \frac{\sum_{\ell=1}^{L}\sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell}{\sum_{\ell=1}^{L}\sum_{i_\ell=1}^{M_\ell} |G_\ell|k_\ell}, \tag{A.63}$$

where M_ℓ and k_ℓ are, respectively, the number of moduli and the number of templates at layer ℓ, and G_ℓ is the pooling range at each layer of the hierarchical architecture.

Proof: The first layer is by construction implementing a $\bar{G}_1$ invariance. Thus we need $G_1 \supseteq \bar{G}_1$ transformations per each of the k_1 templates and M_1 modules (assuming the same number of templates at each module),

$$c_1 = \sum_{i_1=1}^{M_1} |G_1|k_1,$$

where M_1 is the number of parts (modules) of the image I we consider at the first layer.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

For the second layer,

$$c_2 = \sum_{i_1=1}^{M_2} |G_2/G_1|k_2.$$

The key point is that to implement the G_2 invariance we need fewer transformations, G_2/G_1, since the representation at the first layer is already G_1 invariant). This is true, since $|\text{supp}(\langle t^k, gt^k \rangle)| = 1$ and therefore $\tilde{G}_1 = G_1$.

In general, at layer ℓ,

$$c_\ell = \sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell, \ \ell = 1, \ldots, L.$$

Summing the contributions at each layer, we have for the hierarchical construction

$$c(L) \propto \sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell. \tag{A.64}$$

For the corresponding multiple single-layer construction we have instead

$$c_\ell = \sum_{i_\ell=1}^{M_\ell} |G_\ell|\tilde{k}_\ell, \ \ell = 1, \ldots, L,$$

where, as demonstrated in the previous section, $\tilde{k}_\ell = k_\ell$. Note that in this case to implement a G_ℓ invariance at layer ℓ we have to use a bigger set of transformations, G_ℓ (versus $G_\ell/G_{\ell-1}$ of the hierarchy), since no invariance is inherited from the layers below. Thus,

$$c(1) = \sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell|k_\ell, \tag{A.65}$$

which gives a ratio

$$r = \frac{c(L)}{c(1)} = \frac{\sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell/G_{\ell-1}|k_\ell}{\sum_{\ell=1}^{L} \sum_{i_\ell=1}^{M_\ell} |G_\ell|k_\ell}. \tag{A.66}$$

The theorem proves that far fewer stored transformations of templates are needed to achieve invariance for a group transformation G with hierarchical architecture than with shallow architecture.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A.11 Nonlinearities Are Key in Gaining Compression from Repeated Random Projections

The selectivity property of the invariant representation proposed here is guaranteed by an adequate number of projections onto random templates (or PCs) (see section A.2). In the following we prove that repeating this projection step does not provide any computational advantage unless some nonlinearities came into play.

A.11.1 Hierarchical (Linear) Johnson-Lindenstrauss Lemma

Let $\{x_1, x_2, \ldots, x_n\}$ be a set of vectors in $\mathbb{R}^d$. The J-L lemma says that for any $0 < \epsilon < 1/2$ there exists an $f : \mathbb{R}^d \to \mathbb{R}^k$ for $k = O(\epsilon^{-2}) \log(n)$ such that

$$\forall\, i,j\ (1-\epsilon) \left\|x_i - x_j\right\|_2 \leq \left\|f(x_i) - f(x_j)\right\|_2 \leq (1+\epsilon) \left\|x_i - x_j\right\|_2 .$$

The J-L lemma is equivalent to the following statement [132].

Lemma A.28 For any $0 < \epsilon < 1/2$ and positive integer d, there exists a distribution $\mathcal{D}$ over $\mathbb{R}^{k \times d}$ for $k = O(\epsilon^{-2} log(1/\delta))$ such that for any $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$,

$$Pr_{A \sim \mathcal{D}}[\,|\, \|Ax\|_2^2 - 1| > \epsilon] < \delta.$$

We want to formulate first a local version of the J-L lemma and then a recursive/hierarchical one. For the local version suppose $x \in \mathbb{R}^{d \times p}$. Let $A = \mathbb{I}_p \otimes A$ and $x_k = P_k x$ be the projection onto the kth block of d-components of the x vector, $k = 1, \ldots, p$. If we suppose $Pr_{A \sim \mathcal{D}}[\,|\, \|Ax_k\|_2^2 - 1| > \epsilon] < \delta \ \ \forall k$, then also

$$Pr_{A \sim \mathcal{D}}[\,|\, \left\|(\mathbb{I}_p \otimes A)x\right\|_2^2 - 1| > \epsilon] < \delta.$$

Therefore instead of applying the J-L lemma to the whole vector x, we can write it as $x = x_1 \oplus x_2 \oplus \ldots \oplus x_p$ and apply it to each of the p parts, and the embedding still holds. This result says that there is no difference between the applying J-L lemma locally or globally. We can then repeat the procedure on the resulting vector $y \in \mathbb{R}^{k \times p}$. However, the number of different vectors obtained after the projection will be still n, and consequently we will need again the same number of projections, so no gain.

Consider now a different situation in which the number of images depends on the vector size. In this case, we will have an effective gain in repeating the J-L projections.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

A.11.2 Example of Nonlinear Extension of Johnson-Lindenstrauss Lemma

Suppose we have an architecture composed of two-layers; the first layer is composed of P parts, each of d pixels, and the second layer of one part of P pixels. Suppose also we have a one-layer architecture composed of Pd pixels. We want to compare the number of templates needed to discriminate among the different possible images that are inputs to the two architectures. The reasoning holds for any choice of a dictionary. We choose a PC dictionary as a possible example.

In order to use the J-L result we have to know the number of possible images for a certain image size s. We can say that they correspond to the number of all possible combinations of the pixels at different (finite) gray levels. Here, however, we adopt a different criterion. Suppose we have N images of size s (N can be as big as we want and in the limit goes to infinity). The criterion we adopt is the following. The number of possible images of size s is the number of images we can distinguish (up to an ϵ factor) and is given by the number of possible images we can build using a number of PCs (extracted from the N images of size s, $N \to \infty$) such that they allow a reconstruction up to an ϵ error; we call this number $f(s)$. If we further suppose that the PCs are linearly combined with coefficients that can take a finite (but arbitrarily large) number of values, say V, we have that the number of possible images with size s is given by

$$V^{f(s)}.$$

Let us now go back to the one-layer versus the two-layer comparison. If we assume that according to the J-L lemma the number of templates is given (up to a constant) by the log of the number of possible images, we have that the number of templates for the one-layer architecture is

$$\ln(V^{f(Pd)})$$

and for the two-layers architecture

$$P\ln(V^{f(d)}) + \ln(V^{f(P)}).$$

Suppose we can reuse the template projections for the different parts. Given that the log and the power are monotonically increasing functions, we are left with the comparison

$$f(d) + f(P) \lesseqgtr f(dP).$$

From a look at the plot of the number of PCs versus eigenvalues values at different sizes (see [9], figure 3), we have

$$\lambda(i,s) \approx e^{-\frac{i}{s}},$$

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

where the i axis is the count of the number of PCs. Reconstruction up to an ϵ factor will correspond to a constant $T(\epsilon)$ such that $\lambda < T(\epsilon)$, that is, with a simple inversion of the formula we have that the number of PCs that allows a reconstruction with an error up to ϵ is

$$f(s) = -\text{Round}\ln(T(\epsilon))s).$$

Neglecting for the moment the round approximation, we have that the comparison between the one- and two-layer architecture boils down to

$$\text{(two layers)} \quad d + P \lesseqgtr Pd \quad \text{(one layer)}.$$

For $P > 1$ (more then one part) we have that $d + P < Pd$ is always true, that is, there are fewer templates for the two-layer architecture than for the one-layer architecture. Any choice of dictionary that allows the same reconstruction error with fewer dictionary elements will give a similar result. The underlying idea is that images have structure and can be expressed with a dictionary of cardinality much lower than the number of pixels.

Remark A.29 This is an application of the J-L lemma in a nonlinear setting. In a hierarchical architecture the nonlinearity is the indicator function on the parts.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

References

[1] F. Anselmi, L. Rosasco, and T. Poggio. On invariance and selectivity in representation learning. arXiv:1503.05938; Center for Brains, Minds and Machines, CBMM Memo No. 029, 2015.

[2] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* 36 (1980): 193–202.

[3] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2 (1999): 1019–1025.

[4] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. arXiv:1311.2901, 2013.

[5] M. C. Potter. Meaning in visual search. *Science* 187, no. 4180 (1975): 965–966.

[6] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature* 381, no. 6582 (1996): 520–522.

[7] Q. Liao, J. Z. Leibo, and T. Poggio. Learning invariant representations and applications to face verification. In *Proceedings of the 2013 Conference on Neural Information Processing Systems.*

[8] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio. Unsupervised learning of invariant representations in hierarchical architectures. arXiv:1311.4158, 2014: *Theoretical Computer Science,* in press, 2015.

[9] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio. Magic materials: A theory of deep hierarchical architectures for learning sensory representations. Center for Biological and Computational Learning, MIT, 2013.

[10] F. Anselmi and T. Poggio. Representation learning in sensory cortex: A theory. Center for Brains, Minds and Machines, CBMM Memo No. 026, 2014.

[11] T. Poggio, J. Mutch, F. Anselmi, A. Tacchetti, L. Rosasco, and J. Z. Leibo. Does invariant recognition predict tuning of neurons in sensory cortex?. Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, Technical Report 019, 2013; Center for Biological and Computational Learning (CBCL), MIT, Paper No. 313, 2013.

[12] F. Anselmi, L. Rosasco, T. Cheston, and T. Poggio. Deep convolutional networks are hierarchical kernel machines. Center for Brains, Minds and Machines, CBMM Memo No. 035, 2015.

[13] J. Z. Leibo, Q. Liao, F. Anselmi, and T. Poggio. The invariance hypothesis implies domain-specific regions in visual cortex. bioRxiv:10.1101/004473, 2014.

[14] T. Poggio, J. Mutch, and L. Isik. Computational role of eccentricity dependent cortical magnification. arXiv:1406.1770; Center for Brains, Minds and Machines, CBMM Memo No. 017, 2014.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

[15] J. Z. Leibo, F. Anselmi, J. Mutch, A. F. Ebihara, W. Freiwald, and T. Poggio. View-invariance and mirror-symmetric tuning in a model of the macaque face processing system. In *Proceedings of the 2013 Conference on Computational and Systems Neuroscience,* poster 1–54.

[16] T. Poggio and S. Smale. The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society* 50, no. 5 (2003): 537–544.

[17] T. Lee and S. Soatto. Video-based descriptors for object recognition. *Image and Vision Computing* 29 (2012): 639–652.

[18] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. In *Proceedings of the IEEE* 86 (1998): 2196–2209.

[19] Y. Abu-Mostafa. Hints and the VC dimension. *Neural Computation* 5 (1993): 278–288.

[20] D. Hubel and T. Wiesel. Receptive fields: Binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160, no. 1 (1962): 106.

[21] S. Chikkerur and T. Poggio. Approximations in the HMAX model. Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, Technical Report 021, 2011.

[22] M. Kouh and T. Poggio. A canonical neural circuit for cortical nonlinear operations. *Neural Computation* 20, no. 6 (2008): 1427–1451.

[23] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America* 2, no. 2 (1985): 284–299.

[24] D. Gabor. Uncertainty principles and signal recovery. *Journal of the Institution of Electrical Engineers. Part III: Radio and Communication Engineering* 93, no. 26 (1946): 429–441.

[25] D. L. Donoho and S. P. Stark. Uncertainty principles and signal recovery. *SIAM Journal* on *Applied Mathematics* 49, no. 3 (1989): 906–931.

[26] D. Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology* 88, no. 1 (2002): 455–463.

[27] J. P. Jones and L. A. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology* 58, no. 6 (1987): 1233–1258.

[28] C. Niell and M. Stryker. Highly selective receptive fields in mouse visual cortex. *Journal of Neuroscience* 28, no. 30 (2008): 7520–7536.

[29] H. Bülthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences* 89, no. 1 (1992): 60–64.

[30] M. J. Tarr. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three dimensional objects. *Psychonomic Bulletin and Review* 2, no. 1 (1995): 55–82.

[31] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, no. 6583 (1996): 607–609.

[32] S. Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics* 65, no. 10 (2012): 1331–1398.

[33] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience* 3 (2000): 1199–1204.

[34] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1, no. 4 (1989): 541–551.

[35] J. Z. Leibo, J. Mutch, and T. Poggio. Why the brain separates face recognition from object recognition. In *Proceedings of the 2011 Conference on Neural Information Processing Systems*.

[36] D. Arathorn. Computation in the higher visual cortices: Map-seeking circuit theory and application to machine vision. In *Proceedings of the 2004 International Symposium on Information Theory,* 73–78.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

[37] L. Sifre, S. Mallat, and P. France. Combined scattering for rotation invariant texture analysis. In *Proceedings of the 2012 European Symposium on Artificial Neural Networks,* 127–132.

[38] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Center for Biological and Computational Learning (CBCL), MIT, Paper No. 259, 2005.

[39] S. S. Chikkerur, T. Serre, C. Tan, and T. Poggio. What and where: A Bayesian inference theory of attention. *Vision Research* 50 (2010): 2233–2247.

[40] D. George and J. Hawkins. A hierarchical Bayesian model of invariant pattern recognition in the visual cortex. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks,* 1812–1817.

[41] S. Geman. Invariance and selectivity in the ventral visual pathway. *Journal of Physiology (Paris)* 100, no. 4 (2006): 212–224.

[42] S. Hochstein and M. Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron* 36, no. 5 (2002): 791–804.

[43] H. Schulz-Mirbach. Constructing invariant features by averaging techniques. In *Proceedings of the 1994 International Conference on Pattern Recognition. Conference B,* 387–390.

[44] B. Haasdonk and H. Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine Learning* 68 (July 2007): 35–61.

[45] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G.R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* 11 (2010): 1517–1561.

[46] R. Rosenholtz. General-purpose localization of textured image regions. In *Proceedings of the 1999 SPIE Conference on Human Vision and Electronic Imaging,* 4: 454–460.

[47] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition. *Nature Neuroscience* 2 (1999): 1019–1025.

[48] J. Freeman and E. Simoncelli. Metamers of the ventral stream. *Nature Neuroscience* 14 (2011): 1195–1201.

[49] P. Földiák. Learning invariance from transformation sequences. *Neural Computation* 3, no. 2 (1991): 194–200.

[50] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences* 104, no. 15 (2007): 6424–6429.

[51] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. In *The Handbook of Brain Theory and Neural Networks,* ed. M. A. Arbib, 276–278. 2nd ed. Cambridge, MA: MIT Press, 2002.

[52] T. Poggio. A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology* 55 (1990): 899–910.

[53] D. Perrett and M. Oram. Neurophysiology of shape processing. *Image and Vision Computing* 11, no. 6 (1993): 317–333.

[54] B. W. Mel. SEEMORE: Combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation* 9 (1997): 777–804.

[55] S. M. Stringer and E. T. Rolls. Invariant object recognition in the visual system with novel views of 3D objects. *Neural Computation* 14 (2002): 2585–2596.

[56] N. Pinto, J. J. DiCarlo, and D. Cox. How far can you get with a modern face recognition test set using only simple features? In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition,* 2591–2598.

[57] A. M. Saxe, M. Bhand, R. Mudur, B. Suresh, and A. Y. Ng. Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In *Proceedings of the 2011 Conference on Neural Information Processing Systems,* 1971–1979.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

[58] Q. V. Le, R. Monga, M. Devin, G. Corrado, K. Chen, M. Ranzato, J. Dean, and A. Y. Ng. Building high-level features using large-scale unsupervised learning. arXiv:1112.6209, 2011.

[59] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 4277–4280.

[60] D. O. Hebb. *The Organization of Behavior: A Neuropsychological Theory.* New York: Wiley, 1949; new edition published 2002 by Lawrence Erlbaum Associates, Mahwah, NJ.

[61] G. G. Turrigiano and S. B. Nelson. Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience* 5, no. 2 (2004): 97–107.

[62] E. Oja. Simplied neuron model as a principal component analyzer. *Journal of Mathematical Biology* 15, no. 3 (1982): 267–273.

[63] J. Karhunen. Stability of Oja's PCA subspace rule. *Neural Computation* 6 (1994): 739–747.

[64] T. Sanger. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks* 2, no. 6 (1989): 459–473.

[65] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks* 5, no. 6 (1992): 927–935.

[66] A. Hyvrinen and E. Oja. Independent component analysis by general nonlinear Hebbian-like learning rules. *Signal Processing* 64 (1998): 301–313.

[67] D. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems* 5 (1994): 517–548.

[68] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems* 14 (2003): 391–412.

[69] J. Gallant, J. Braun, and D. V. Essen. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science* 259 (1993): 100–103.

[70] J. Gallant, C. Connor, S. Rakshit, J. Lewis, and D. Van Essen. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of Neurophysiology* 76 (1996): 2718–2739.

[71] J. Hegde and D. Van Essen. Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience* 20, no. 5 (2000): 61.

[72] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way.* 3rd ed. Burlington, MA: Academic Press, 2008.

[73] D. Marr, T. Poggio, and E. Hildreth. Smallest channel in early human vision. *Journal of the Optical Society of America* 70, no. 7 (1980): 868–870.

[74] A. Cowey and E. T. Rolls. Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research* 21, no. 5 (1974): 447–454.

[75] H. Strasburger, I. Rentschler, and M. Jüttner. Peripheral vision and pattern recognition: A review. *Journal of Vision* 11, no. 5 (2011): 13.

[76] S. Ullmon. Object recognition and segmentation by a fragment-based hierarchy. Trends in Cognitize Sciences 11 (2); 58–69, 2007.

[77] J. A. Mazer, W. E. Vinje, J. McDermott, P. H. Schiller, and J. L. Gallant. Spatial frequency and orientation tuning dynamics in area V1. *Proceedings of the National Academy of Sciences* 99, no. 3 (2002): 1645–1650.

[78] D. Hubel and T. Wiesel. Uniformity of monkey striate cortex: A parallel relationship between field size, scatter, and magnification factor. *Journal of Comparative Neurology* 158 (1974): 295–305.

[79] R. Gattass, C. Gross, and J. Sandell. Visual topography of V2 in the macaque. *Journal of Comparative Neurology* 201, no. 4 (1981): 519–539.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

[80] R. Gattass, A. Sousa, and C. Gross. Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience* 8, no. 6 (1988): 1831–1845.

[81] D. Levi and T. Carney. The effect of flankers on three tasks in central, peripheral, and amblyopic vision. *Journal of Vision* 11, no. 1 (2011): 10.

[82] H. Bouma. Interaction effects in parafoveal letter recognition. *Nature* 226 (1970): 177–178.

[83] D. Pelli and K. Tillman. The uncrowded window of object recognition. *Nature Neuroscience* 11, no. 10 (2008): 1129–1135.

[84] D. Levi. Crowding, an essential bottleneck for object recognition: A minireview. *Vision Research* 48, no. 5 (2008): 635–654.

[85] M. Lev, O. Yehezkel, and U. Polat. Uncovering foveal crowding? *Scientific Reports* 4 (2014): article 4067.

[86] P. G. Schyns and F. Gosselin. Diagnostic use of scale information for componential and holistic recognition. In *Perception of Faces, Objects, and Scenes: Analytic and Holistic Processes,* ed. M. A. Peterson and G. Rhodes, 120–145. New York: Oxford University Press, 2003.

[87] S. Anstis. A chart demonstrating variations in acuity with retinal position. *Vision Research* 14, no. 7 (1974): 589–592.

[88] H. D. Beeck and R. Vogels. Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology* 426, no. 4 (2000): 505–518.

[89] J. DiCarlo and J. Maunsell. Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology* 89, no. 6 (2003): 3264–3278.

[90] B. Balas, L. Nakano, and R. Rosenholtz. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision* 9, no. 12 (2009):13.

[91] D. Tsao and W. Freiwald. Faces and objects in macaque cerebral cortex. *Nature* 6, no. 9 (2003): 989–995.

[92] J. Z. Leibo, J. Mutch, and T. Poggio. How can cells in the anterior medial face patch be viewpoint invariant? Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, Technical Report 057, 2010.

[93] N. Kanwisher. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences* 25, no. 107 (2010): 11163–11170.

[94] P. Downing and Y. Jiang. A cortical area selective for visual processing of the human body. *Science* 293, no. 5539 (2001): 2470–2073.

[95] M. Bar, E. Aminoff, and D. L. Schacter. Scenes unseen: The parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. *Journal of Neuroscience* 28, no. 34: 8539–8544.

[96] N. Logothetis, J. Pauls, H. Bülthoff, and T. Poggio. View-dependent object recognition by monkeys. *Current Biology* 4, no. 5. (1994): 401–414.

[97] N. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology* 5, no. 5 (1995): 552–563.

[98] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature* 343 (1990): 263–266.

[99] S. Ku, A. Tolias, N. Logothetis, and J. Goense. fMRI of the face-processing network in the ventral temporal lobe of awake and anesthetized macaques. *Neuron* 70, no. 2 (2011): 352–362.

[100] R. Malach, J. B. Reppas, R. R. Benson, K. Kwong, H. Jiang, W. A.Kennedy, P. J. Ledden, T. J. Brady, B. R. Rosen, and R. B. Tootell. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences* 92, no. 18 (1995): 8135–8139.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

[101] L. Cohen, S. Dehaene, L. Naccache, S. Lehéricy, G. Dehaene-Lambertz, M. Hénaff, and F. Michel. The visual word form area. *Brain* 123, no. 2 (2000): 291–307.

[102] N. Kriegeskorte. Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2, no. 4 (2008).

[103] N. Logothetis and D. Sheinberg. Visual object recognition. *Annual Review of Neuroscience* 19, no. 1 (1996): 577–621.

[104] M. Stryker. Temporal associations. *Nature* 354 (1991): 108–109.

[105] W. Freiwald and D. Tsao. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* 330, no. 6005 (2010): 845–851.

[106] W. Freiwald. A face feature space in the macaque temporal lobe. *Nature Neuroscience* 12 (2009): 1187–1196.

[107] J. Z. Leibo. The invariance hypothesis and the ventral stream. Ph.D. diss., Massachusetts Institute of Technology, 2013.

[108] J. Mutch, J. Z. Leibo, S. Smale, L. Rosasco, and T. Poggio. Neurons that confuse mirror-symmetric object views. Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT, Technical Report 062, 2010; Center for Biological and Computational Learning (CBCL), MIT, Paper No. 295, 2010.

[109] Y. LeCun. LeNet-5, Convolutional Neural Networks. http://yann.lecun.com/exdb/lenet /index.html. Accessed Nov. 2015.

[110] B. C. Motter and G. F. Poggio. Blmoculer fixation in the rhesus monkey: spatial and temporal characteristics. Exp Brain Res. 54, 304–314, 1984.

[111] E. Bricolo. On the representation of novel objects: Human psychophysics, monkey physiology, and computational models. Ph.D. diss., Massachusetts Institute of Technology, 1996.

[112] D. Tsao and M. Livingstone. Mechanisms of face perception. *Annual Review of Neuroscience* 31 (2008): 411–437.

[113] C. P. Hung, G. Kreiman, T. Poggio, and J. J. DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310 (2005): 863–866.

[114] A. Hyvrinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks* 13 (2000): 411–430.

[115] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, no. 12 (2008): 2109–2125.

[116] D. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience* 9 (1992):181–197.

[117] T. Poggio and W. Reichardt. Considerations on models of movement detection. *Kybernetik* 13 (1973): 223–227.

[118] J. Mutch and D. Lowe. Multiclass object recognition with sparse localized features. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition,* 11–18.

[119] Y. LeCun, F. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition,* 2: 97–104.

[120] G. Palm. On associative memory. *Biological Cybernetics* 36, no. 1 (1980): 19–31.

[121] T. Poggio. On optimal nonlinear associative recall. *Biological Cybernetics* 19, no. 4 (1975): 201–209.

[122] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society* 39, no. 1 (2001): 1–49.

[123] H. Cramér and H. Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society* 4 (1936): 290–294.

[124] J. Cuesta-Albertos, R. Fraiman, and T. Ransford. A sharp form of the Cramér-Wold theorem. *Journal of Theoretical Probability* 20, no. 2 (2007): 201–209.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

[125] J. Cuesta-Albertos. How many random projections suffice to determine a probability distribution? IPMs sections, 2009.

[126] A. Heppes. On the determination of probability distributions of more dimensions by their projections. *Acta Mathematica Hungarica* 7, no. 3 (1956): 403–410.

[127] T. Plate. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence,* 30–35.

[128] L. Isik, E. M. Meyers, J. Z. Leibo, and T. Poggio. The timing of invariant object recognition in the human visual system. *Journal of Neurophysiology* 111, no. 1 (2013): 91–102.

[129] J. Antoine, R. Murenzi, P. Vandergheynst, and S. Ali. *Two-Dimensional Wavelets and Their Relatives.* Cambridge: Cambridge University Press, 2004.

[130] C. F. Stevens. Preserving properties of object shape by computations in primary visual cortex. *Proceedings of the National Academy of Sciences* 101, no. 11 (2004): 15524–15529.

[131] A. Grossman, J. Morlet, and T. Paul. Transforms associated to square integrable group representations: Examples. *Annales de l'I.H.P Probabilités et statistiques* 45 (1986): 293–309.

[132] A. Saxe, P. W. Koh, Z. Chen, M. Bhand, B. Suresh, and A. Ng. On random weights and unsupervised feature learning. In *Proceedings of the 2011 International Conference on Machine Learning,* 1089–1096.

[133] J. Nelson. Johnson-Lindenstrauss notes. http://web.mit.edu/minilek/www/jl_notes.pdf. Accessed Nov. 2015.

[134] Center for Brains, Minds and Machines. CBMM Turing++ questions. http://cbmm.mit.edu/research/cbmm-challenge. Accessed Nov. 2015.

[135] H.J.W. Müller-Kirsten. *Introduction to Quantum Mechanics.* Singapore: World Scientific, 2006.

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Index

Abelian group, 68, 100
  non-Abelian, 67, 102, 104

Backprojection, xii, 60, 62
Bouma's law, 36, 38–39, 42

Categorization, 1, 16, 56. *See also* Identification
Class-specific, xiii, 11, 13, 15, 45–51, 53, 56, 58
Clutter, 35, 37, 41–42, 60–61, 79
Complex cell, xi, 5, 8–10, 14, 19, 20, 23–25, 32, 37–38, 42–43, 55, 58–59, 62, 83–84, 86
  and pooling, 23, 36, 37
  wiring, 24, 86
Convolutional networks, 14, 21, 54, 57
  group convolution, 82
Cortical equation, 87–92
Covariance, 15, 19–20, 25, 67, 80, 82–84, 87, 100
  matrix, operator, 26–28, 49, 88
  and wavelets, 84
Cramér-Wold theorem, 69–70, 72. *See also* Johnson-Lindenstrauss lemma
Crowding, 36, 38–39, 42

Decimation, 36
Deep convolutional networks. *See* Convolutional networks

Eccentricity, xiii, 25, 29, 30–32, 35, 37–42, 54–55, 58

Face patch, xiii, 4, 13, 48–50, 58, 86
Factorization of invariances, 15, 19–20, 98–101
Feedback, 59, 60

Fovea, 30–33, 38–42
Foveola, 32–33, 36–42

Gabor function, xiii, 10–11, 14, 30–31, 77–78
Genes 29, 57
Group transformations, xii, 5–7, 14, 27–28, 54, 57, 72, 74. *See also* Nongroup transformations
  affine, xiii, 7, 45, 67, 86, 92, 94
  approximation, 19
  average, 8, 54, 72
  convolution, 82
  Euclidean, 100
  factorization, 19–20, 98–100
  Lie, 97
  locally compact, xiii, 9–10, 68, 76, 80, 87, 94, 104
  partially, observable, 9, 75
  sparsity, 79
  subgroup, 98
  translation, 5, 57–58, 66, 87

Hebbian learning, xiii, 24–25, 35, 88
Heisenberg principle, 78, 94
Hierarchical architectures, 14–15, 17–20, 42, 61, 84–85, 98, 101–105
HMAX, xi, 14, 23, 25, 54, 61–62, 71
Hubel and Wiesel, xi, 3, 14, 32, 57

ICA (independent component analysis), 59
Identification, 1, 17, 26
Inferotemporal cortex (IT), 56
Invariance
  class-specific, 11, 13, 45
  cost, 104
  and factorization, 98
  and group average, 8
  and localization, 14, 75–80, 84, 94–97
  and nongroup, xii, 3, 7, 11, 13, 19, 58, 79

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Invariance (cont.)
optimal, 54, 58, 77–79
quasi invariance, xiii, 2, 11, 58, 79
representation, xii, 1–10
and sample complexity, xii, 1–5, 15, 65–68
scale invariance, xiii, 10, 27, 31, 39–41, 77, 95–96
signature, 5–6, 14, 20, 28
translation invariance, 4, 27, 39, 94, 96

Johnson-Lindenstrauss lemma, 73, 103, 106–107. *See also* Random projections

Lie group transformations, 97. *See also* Group transformations

Memory, 2, 15–16, 17, 36, 40, 42, 56
associative, 60, 62–63
Mirror symmetry, 49–51
Multiresolution, xiii, 32, 58

Nongroup transformations, xii, 3, 7, 11, 13, 19, 58, 79. *See also* Quasi invariance

Object recognition, xi, 1, 18, 53. *See also* Categorization
Occlusion, 7, 18
Oja's rule, flow, 25–26, 27, 59. *See also* Hebbian learning
Optimal templates xiii, 10, 31, 70. *See also* Gabor function

Partially observable group 9, 10, 19, 78, 80. *See also* Group transformations
PIT (posterior inferior temporal cortex), 3, 36, 37, 45, 46
Plasticity, 26, 28, 56, 59. *See also* Hebbian learning
Pooling, xi, xiii, 5, 19, 21–22, 30, 40, 41, 47, 59
and clutter, 35
complex cell, 23, 24, 26, 86
and deep networks, 54–55
inverted truncated pyramid, 37–38
multistage, 35
PCA, 27–28, 49–51, 86
range, 13, 15, 20, 84–85, 94, 97, 103–104
scale translation, 34, 35
size, 72
Principal components, 24, 26, 43, 49, 87
and mirror symmetry 49–51

Quasi invariance, xiii, 2, 79. *See also* Nongroup transformations

Random projections, 69–70, 73, 103, 106. *See also* Cramér-Wold theorem; Johnson-Lindenstrauss lemma
Retina, xiii, 20, 35, 39–40, 58

Sample complexity, xii, 1, 2, 4–5, 15, 61, 65–68
Scattering transform, 14, 22, 58
Selectivity, 4, 7, 8, 24, 27, 68, 80, 106
Signature, xii, 5–6, 9, 14, 15, 18, 20, 41, 42, 46, 53, 56, 61, 62, 69, 70, 72, 75, 81, 83, 86, 97, 98, 99, 101
local, 9, 16, 60
global, 16, 85
and PCA, 28
recursive, 80
Simple cell, 5, 11, 14, 16, 19, 29, 32, 38, 39, 41, 50, 55, 59, 62, 80, 81, 84, 97. *See also* Complex cell
and class specificity, 45
and deep networks, 22
and genes, 57
tuning, 23–28, 35, 42 (*see also* Gabor function)
Sparsity, 12, 14, 79
and incoherence, 12, 13, 14
and localization, 14
Supervised learning, xii, 1, 4, 47, 48, 58, 61. *See also* Unsupervised learning

Truncated inverted pyramid, 30, 31, 32, 33, 39, 40, 42, 58
multistage, 37
Tuning, xiii, 13, 23, 24, 25, 28, 35, 42, 45, 55, 56, 61, 63, 88
mirror-symmetric, 49–51
simple cells, 22–28, 35, 42

Unsupervised learning, 5, 6, 23, 56, 60, 61. *See also* Supervised learning

V1, xiii, 3, 11, 18, 25, 26, 27, 28, 29, 35, 36, 37, 38, 39, 40, 51, 54, 55, 56, 57, 63
scale and position invariance, 34
tuning, 35
V2, V4, 35–42
tuning 43
Visual concepts, 63
Visual cortex, 5, 54, 55, 61, 63, 87, 88. *See also* Inferotemporal cortex; PIT

Wavelets, 14, 27, 31, 58, 78, 79
and covariance, 84
and localization, 77, 92

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Computational Neuroscience

Terence J. Sejnowski and Tomaso A. Poggio, editors

Neural Nets in Electic Fish, Walter Heiligenberg, 1991

The Computational Brain, Patricia S. Churchland and Terrence J. Sejnowski, 1992

Dynamic Biological Networks: The Stomatogastic Nervous System, edited by Ronald M. Harris-Warrick, Eve Marder, Allen I. Selverston, and Maurice Moulins, 1992

The Neurobiology of Neural Networks, edited by Daniel Gardner, 1993

Large-Scale Neuronal Theories of the Brain, edited by Christof Koch and Joel L. Davis, 1994

The Theoretical Foundations of Dendritic Function: Selected Papers of Wilfrid Rall with Commentaries, edited by Idan Segev, John Rinzel, and Gordon M. Shepherd, 1995

Models of Information Processing in the Basal Ganglia, edited by James C. Houk, Joel L. Davis, and David G. Beiser, 1995

Spikes: Exploring the Neural Code, Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialck, 1997

Neurons, Networks, and Motor Behavior, edited by Paul S. Stein, Sten Grillner, Allen I. Selverston, and Douglas G. Stuart, 1997

Methods in Neuronal Modeling: From Ions to Networks, second edition, edited by Christof Koch and Idan Segev, 1998

Fundamentals of Neural Netowrk Modeling: Neuropsychology and Cognitive Neuroscience, edited by Randolph W. Parks, Daniel S. Levine, and Debra L. Long, 1998

Neural Codes and Distributed Representations: Foundations of Neural Computation, edited by Laurence Abbott and Terrence J. Sejnowski, 1999

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024

Unsupervised Learning: Foundations of Neural Computation, edited by Geoffrey Hinton and Terrence J. Sejnowski, 1999

Fast Oscillations in Cortical Circuits, Roger D. Traub, John G. R. Jeffreys, and Miles A. Whittington, 1999

Computational Vision: Information Processing in Perception and Visual Behavior, Hanspeter A. Mallot, 2000

Graphical Models: Foundations of Neural Computation, edited by Michael I. Jordan and Terrence J. Sejnowski, 2001

Self-Organizing Map Formation: Foundations of Neural Computation, edited by Klaus Obermayer and Terrence J. Sejnowski, 2001

Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems, Chris Eliasmith and Charles H. Anderson, 2003

The Computational Neurobiology of Reaching and Pointing, Reza Shadmehr and Steven P. Wise, 2005

Dynamical Systems in Neuroscience, Eugene M. Izhikevich, 2006

Bayesian Brian: Probabilistic Approaches to Neural Coding, edited by Kenji Doya, Shin Ishii, Alexandre Pouget, and Rajesh P. N. Rao, 2007

Computational Modeling Methods for Neuroscientists, edited by Erik De Schutter, 2009

Neural Control Engineering, Steven J. Schiff, 2011

Understanding Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging, edited by Nikolaus Kriegeskorte and Gabriel Kreiman, 2011

Biological Learning and Control: How the Brain Builds Representations, Predicts Events, and Makes Decisions, Reza Shadmehr and Sandro Mussa-Ivaldi, 2012

Principles of Brain Dynamics: Global State Interactions, edited by Mikhail Rabinovich, Karl J. Friston, and Pablo Varona, 2012

Brain Computation as Hierarchical Abstraction, Dana H. Ballard, 2015

Visual Cortex and Deep Networks: Learning Invariant Representations, Tomaso A. Poggio and Fabio Anselmi, 2016

Case Studies in Neural Data Analysis: A Guide for the Practicing Neuroscientist, Mark A. Kramer, and Uri T. Eden, 2016

From Neuron to Cognition via Computational Neuroscience, edited by Michael A. Arbib and James J. Bonajuto, 2016

Downloaded from http://direct.mit.edu/books/book-pdf/2161172/book_9780262336710.pdf by guest on 21 November 2024